Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

A Window into the Infant's Visual World: Through a NeuroAI Lens

Áine Travers Dineen, B.A. (*Mod.*)

School of Psychology and Trinity College Institute of Neuroscience

Trinity College Dublin

A thesis submitted to

Trinity College Dublin, University of Dublin

For the degree of

Doctor of Philosophy

2026

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work. I confirm that the work has been completed in full compliance with relevant university policies, including the Academic Integrity policy, the Trinity Policy on Good Research Practice, and the guidance on AI and GenAI in Teaching, Learning, Assessment and Research.

I agree to deposit this thesis in the University's open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).

Áine Travers Dineen

Summary

For centuries, scientists, philosophers, and caregivers have wondered what infants perceive of the world around them. Recent work has greatly advanced our understanding of both perceptual development and the content of infant visual experience. Much less is known, however, about the features encoded by the developing brain when infants view objects.

NeuroAI approaches, which combine neuroimaging with artificial intelligence (AI) models, are revealing the visual features represented across the adult brain. In particular, these techniques have proven highly effective at characterising the features encoded by the ventral visual stream (VVS), the cortical pathway that supports object processing, by comparing functional MRI (fMRI) responses to feature representations in deep neural networks (DNNs). Despite strong interest in the mechanisms of early object learning from both developmental scientists and computer scientists, comparable work has not previously been possible in infants due to the methodological challenges of acquiring suitable data from this cohort. In this thesis, I address this gap using awake infant fMRI and developmentally inspired DNNs to characterise the spatial scale of the features encoded in infant visual cortex.

In **Chapter 1**, I outline the promise of a NeuroAI approach for characterising developing visual representations and situate this work within the broader literature. In particular, I highlight computational and theoretical accounts proposing that the transient perceptual limitations of infancy may serve as a scaffold for early learning.

In **Chapter 2**, I present an overview of the Foundations of Cognition (FOUNDCOG) project. FOUNDCOG is a longitudinal neurodevelopmental study (N = 134) using awake infant fMRI to characterise developing visual responses at 2 and 9 months of age. I describe the design of two passive-viewing paradigms optimised for infant fMRI: a videos task designed to maximise engagement, and a pictures task designed to enable comparison with computational models. I report acquisition success rates and data-quality metrics, showing that, despite its challenges, awake infant fMRI can yield robust group-level object responses suitable for computational modelling.

In **Chapter 3**, I compare developing VVS responses to developmentally inspired DNNs trained on object images with graded amounts of blur to systematically vary sensitivity to coarser and finer object features. This analysis reveals that V1 already represents both coarse and fine features by 2 months, demonstrating an adult-like pattern of tuning, whereas downstream visual regions show more protracted development and tuning remains comparatively immature at 9 months. These findings suggest that representations in developing V1 are richer and more mature than previously appreciated, while tuning in later visual regions continues to develop across infancy and beyond.

In **Chapter 4**, I isolate spatial frequencies using bandpass filtering and track model–brain correspondence across training to examine the learning dynamics of brain–relevant object information across spatial scales. These analyses reveal distinct inductive biases in minds and machines: DNNs rapidly favoured fine visual features, or high spatial frequencies, whereas infant extrastriate cortex showed a more protracted preference

for shape-based, low-spatial-frequency features. This bias persisted beyond what acuity alone would predict, suggesting that inductive biases in the developing brain scaffold early learning over and above the statistics of visual input.

In **Chapter 5**, I reflect on what this work reveals about developing visual cortex and about the strengths and limitations of DNNs as models of developing object processing. These findings place new empirical constraints on theories of how coarse-to-fine object representations develop in the brain, and show how perceptual constraints and cortical inductive biases jointly shape early object representations. More broadly, this thesis demonstrates that NeuroAI can yield novel insights into the cortical mechanisms of visual development, providing a new window into the infant's visual world.

Acknowledgments

Rhodri, it has been such a privilege working with you and being a member of the FOUNDCOG project team. I am forever grateful for this opportunity. Thank you for encouraging methodological complexity and indulging the seemingly infinite digressions along the way (sometimes at the expense of deadlines, but mostly not, by a very fine margin…) that have made this work richer, perhaps a little more complicated, but certainly a lot more interesting than would otherwise have been possible. On a personal level, I am grateful beyond words for your boundless enthusiasm and support.

Clíona, thank you for all of the time and energy that you spent meticulously curating the FOUNDCOG dataset. This thesis would look very different without your many contributions. Thank you for your strong belief in this emerging field and all of its potential, and for your friendship along this journey. Anna T, I am incredibly grateful to you for your mentorship, particularly in the early days of this project, which laid the foundation for this thesis. I am also grateful to Baby Maeve, who is no longer a baby, and Ian, for the fond memories of mornings spent singing '*The Bare Necessities'* in the office. Graham, thank you for providing the clinical expertise that was instrumental to conducting the awake scans, and for all of the interesting conversations along the way (including the very many hours of meetings to define the tags in Chapter 2). Thank you, Amy, Maebh, Aoibhe and Katie for the hundreds of hours spent precisely cropping and tagging these videos. I am incredibly grateful for your contributions. Thank you to Anna K for your many contributions to the blurry vision project while it was still in its infancy. Lorijn, thank you for your ideas and insights. Thank you to Enna, Chiara, Jessica, and Sojo

for your important contributions to data collection. Thank you to Tammy, Marie and Kevin for bringing new energy to the FOUNDCOG project (and new decorations to the office…). I am so excited for all of the innovative directions that you are pursuing with this precious dataset. Thank you to each of you, and also to Cameron, Karolina, Calum, Amy and Clara, for your enthusiasm and commitment to FOUNDCOG. Thank you wholeheartedly to everyone above for dedicating years of your lives to the FOUNDCOG project, without which none of the work in this thesis would have been possible.

I am also grateful to the extended data collection team: Claire, Katie, Josué, and Anisha, for your assistance with acquisition, and to Michelle, Alessandra, Lucy, Niamh, Linda, Siobhán, and Deirdre for your contributions to clinical supervision during infant scans. Thank you also to Hannah and Oscar for your contributions to stimulus design.

Thank you to our funders, the European Research Council (Advanced Grant ERC-2017-ADG, FOUNDCOG, 787981) and Taighde Éireann – Research Ireland (grants GOIPG/2023/2479 and 22/FFP-A/11050), for making this research possible.

To all of the parents and caregivers at the Coombe and the Rotunda who spent precious minutes during your baby's first days talking to us about our research, whether you decided to participate or not, we are so incredibly grateful to you for your time and enthusiasm. In particular, I wish to sincerely thank all of our FOUNDCOG participants and their families who so generously gave of their time to support this project, without whom this research would not have been possible.

Beyond FOUNDCOG, thank you to Alex for teaching us about visual development, and to the Freeze Motion team, Param, Zi, Faizan, Eric, and collaborators Sarah, Chen, Tom, and Yinsong, for your dedication to the future viability of awake infant fMRI. Thank you to everyone in the TCIN community for the many interesting conversations over the years, and a special thank you to Tina for always greeting our participants with a smile.

Beyond Trinity, I am extremely grateful to the FIT'NG Community for greatly enriching my postgraduate experience. In particular, it has been such a privilege to serve as chair of the Trainee Committee. Thank you to all of the committee members, especially Halie, Paige, Emma, Juliette, Emily, and Tristan. It has been so much fun working with you over the past few years.

Beyond the direct contributors to this work, I am grateful to too many people to name.

Finn, you are so special, and it is such a privilege to be your auntie — you inspire me every day. Sarah, thank you for your generosity and fun-loving spirit. Barry, I am so grateful to you for your encouragement to follow this path before any of us knew where it might lead.

Mostly, thank you to my mother, Lucy, for instilling in me your unwavering passion for child development, and the belief that there is nothing in the world more important than enabling children to reach their full potential through this remarkable period of growth and plasticity. For this, and everything else, I am forever grateful. I dedicate this work to you.

List of Publications and Presentations

Publications

- O'Doherty, C., **Dineen, Á.T.**, Truzzi, A., King, G., Zaadnoordijk, L., Harrison, K., D'Arcy, E., White, J., Caldinelli, C., Holloway, T., Kravchenko, A., Diedrichsen, J., Tarrant, A., Byrne, A., Foran, A., Molloy, E., Cusack, R. (*2026*). Infants have rich visual categories in ventrotemporal cortex at 2 months of age. *Nature Neuroscience*. https://doi.org/10.1038/s41593-025-02187-8

- Chin, J., Wyburd, M., Ayzenberg, V., Bayet, L., Bilgic, B., Chen, E. M., Chen, Y., **Dineen, Á.,** Fujita, S., Liu, J., Jun, Y., Camacho, M. C., Zöllei, L. (*2026*). Deep learning in fetal, infant, and toddler neuroimaging research. https://doi.org/10.1016/j.dcn.2026.101680

- Olson, H.A., Camacho, M.C., Abdurokhmonova, G., Ahmad, S., Chen, E.M., Chung, H., Di Lorenzo, R., **Dineen, Á.T.,** Ganz, M., Licandro, R., Magnain, C., Marrus, N., McCormick, S.A., Rutter, T.M., Wagner, L., Woodruff Carr, K., Zöllei, L., Vaughn, K.A., Skak Madsen, K. (2025). Measuring and interpreting individual differences in fetal, infant, and toddler neurodevelopment. *Developmental Cognitive Neuroscience,* 101539. https://doi.org/10.1016/j.dcn.2025.101539

- Margolis, E.T., Nelson, P.M., Fiske, A., Champaud, J.L., Olson, H.A., Gomez, M.J.C., **Dineen, Á.T.,** Bulgarelli, C., Troller-Renfree, S.V., Donald, K.A. and Spann, M.N. (2025). Modality-level obstacles and initiatives to improve representation in fetal, infant, and toddler neuroimaging research samples. *Developmental Cognitive Neuroscience*, 72, 101505. https://doi.org/10.1016/j.dcn.2024.101505

- Bortolotti, L., Pickles, E., **Dineen, A.**, Sakhabalkar, P., Fuchs, P., Lesbats, C., So, P.W. and Grist, J. (2022). "The MRI education course"–A free resource for the community to provide MRI education, across Europe and Iran, established during the COVID-19 pandemic. *Magnetic Resonance in Medicine.* https://doi.org/10.1002/mrm.29500

Manuscripts in preparation

- **Dineen, Á.T.**, O'Doherty, C., Truzzi, A., Kravchenko, A., King, G., Wade, A.R., D'Arcy, E., Tarrant, A., Byrne, A.T., Foran, A., Molloy, E., Cusack, R. (*in preparation*). The development of tuning to coarse and fine visual features in infant visual cortex.

- **Dineen, Á.T.**, O'Doherty, C., Truzzi, A., King, G., Zaadnoordijk, L., D'Arcy, E.-L., Caldinelli, C., Kravchenko, A., Holloway, T., White, J., Wade, A.R., De Burca, A., Healy, M., Burke, A., Campion, K., Murphy, C., Herbert, K., Ambre, C., Wadhwa, A., Markey, O., Rico-Picó, J., O'Hara, D., Byrne, A.T., Tarrant, A., Foran, A., Molloy, E.J., Cusack, R. (*under review*). Awake infant functional MRI for visual neuroscience: reflections from the Foundations of Cognition project.

- King, G., **Dineen, Á.T.**, O'Doherty, C., Truzzi, A., Caldinelli, C., D'Arcy, E., White, J., Kravchenko, A., Zaadnoordijk, L., Burke, A., Healy, M., Joseph, S., Molloy, E., Foran, A., Tarrant, A., Byrne, A., Cusack, R. (*in preparation*). Practicalities of scanning an awake infant functional MRI cohort.

- Santillo, M. F., O'Doherty, C., **Dineen, Á.T.**, Truzzi, A., King, G., Zaadnoordijk, L., D'Arcy, E., Holloway, T., White, J., Caldinelli, C., Kravchenko, A., Tarrant, A., Byrne, A.T., Foran, A., Molloy, E.J., Cusack, R. (*in preparation*). The origins of memorability in infancy.

Conference proceedings

- Wang, Y., Fletcher, T., Luo, X., **Dineen, Á. T.**, Cusack, R., Qin, C. (2026). Self-supervised slice-to-volume reconstruction with Gaussian representations for fetal MRI. *International Symposium on Biomedical Imaging* (ISBI), London (April 8–11, 2026).

- O'Doherty, C., **Dineen, Á. T.**, Truzzi, A., King, G., D'Arcy, E., Caldinelli, C., Holloway, T., Molloy, E., Cusack, R. (2025). Models trained on infant views are more

predictive of infant visual cortex. *Cognitive Computational Neuroscience* (CCN), Amsterdam.

- O'Doherty, C., **Dineen, Á. T.**, Truzzi, A., King, G., Zaadnoordijk, L., Harrison, K., D'Arcy, E., White, J., Caldinelli, C., Holloway, T., Kravchenko, A., Tarrant, A., Byrne, A., Foran, A., Molloy, E., Cusack, R. (2024). Deep Neural Networks Models of Infant Visual Cortex, *Cognitive Computational Neuroscience* (CCN), Boston.

Invited talks

- **Dineen, Á.T.**, O'Doherty, C., Truzzi, A., King, G., Zaadnoordijk, L., D'Arcy, E.-L., Caldinelli, C., Kravchenko, A., Holloway, T., White, J., Wade, A.R., De Burca, A., Healy, M., Burke, A., Campion, K., Murphy, C., Herbert, K., Ambre, C., Wadhwa, A., Markey, O., Craddock, H., Rico-Picó, J., O'Hara, D., Joseph, S., Byrne, A.T., Tarrant, A., Foran, A., Molloy, E.J., Cusack, R. (2026). Awake infant fMRI for visual neuroscience: reflections from the FOUNDCOG project. Manning Lab Meeting, SickKids, 26 Mar, Toronto, ON.

- **Dineen, Á.T**., O'Doherty, C., Truzzi, A., Kravchenko, A., Wade, A.R., D'Arcy, E.-L., King, G., Foran, A., Molloy, E., Cusack, R. (2026). Development of tuning to coarse and fine visual features: inductive biases differ in minds and machines. Brain and Mind Institute Seminar, Western University, 25 Mar, London, ON.

- **Dineen, Á.T**., O'Doherty, C., Truzzi, A., Kravchenko, A., Wade, A.R., D'Arcy, E.-L., King, G., Foran, A., Molloy, E., Cusack, R. (2026). NeuroAI reveals the development of tuning to coarse and fine visual features in infancy. Turk-Browne Lab Meeting, Yale University, 23 Mar, New Haven, CT.

- **Dineen, Á.T.,** O'Doherty, C., Truzzi, A., Kravchenko, A., King, G., Foran, A., Molloy, E., Cusack, R. (2026). Development of tuning to coarse and fine visual features in infant visual cortex. Colloquium at the Smith-Kettlewell Eye Research Institute (SKERI), 19 Mar, San Francisco, CA.

- **Dineen, Á.T.*,** O'Doherty, C.*, Cusack, R. (2025). The Foundations of Cognition project: using awake infant fMRI to study the development of visual representations in the ventral visual stream. SinhaLab, MIT, 3 Oct, remote presentation. * Denotes equal contribution.

- **Dineen, Á.T.,** O'Doherty, C., Truzzi, A., Kravchenko, A., King, G., Wade, A.R., Zaadnoordijk, L., D'Arcy, E., Holloway, T., White, J., Caldinelli, C., Tarrant, A., Byrne, A.T., Foran, A., Molloy, E., Cusack, R. (2025). The infant ventral visual stream shows hierarchical development of tuning to object statistics. *Post-FIT'NG Workshop*, 10 Sep, Connemara, Galway, Ireland.

- **Dineen, Á.T.,** O'Doherty, C., Truzzi, A., Kravchenko, A., King, G., Foran, A., Molloy, E., Cusack, R. (2025). Characterising the development of object feature tuning in the infant ventral visual stream with deep neural networks. *From Child to Machine Learning: The potential and challenges of translating Developmental Principles into Neural Network Design, Satellite Event at Cognitive Computational Neuroscience* (CCN), 11 Aug, Amsterdam, Netherlands.

- **Dineen, Á.T.,** King, G., O'Doherty, C., Truzzi, A., Caldinelli, C., D'Arcy, E., White, J., Kravchenko, A., Ambre, C., Wadhwa, A., Healy, M., Burke, A., Joseph, S., Molloy, E., Foran, A., Tarrant, A., Byrne, A.T., Cusack, R. (2025). The Foundations of Cognition project: Overview of methods and data acquired. Brain Maturation Group Lab Meeting, Danish Research Centre for Magnetic Resonance, 29 Jun, Copenhagen, Denmark.

- **Dineen, Á.T.,** O'Doherty, C., Truzzi, A., Kravchenko, A., King, G., Wade, A.R., Zaadnoordijk, L., D'Arcy, E., White, J., Holloway, T., Caldinelli, C., Tarrant, A., Byrne, A.T., Foran, A., Molloy, E., Cusack, R. (2024). Blurry beginnings: the infant ventral visual stream emphasises low spatial frequency features. Scaffolding of Cognition Team, Stanford, 20 Dec, Palo Alto, CA.

- O'Doherty, C.*, **Dineen, Á.T.*** (2024). Workshop: applications of artificial intelligence for understanding brain function. *Predictive Brain Health Modelling Group***,** 6 Nov, Trinity College Dublin. * Denotes equal contribution.

- **Dineen, Á.T.**, Cusack, R. (2023). 100 babies: insights from awake fMRI at 2-months. Symposium Contribution, *Flux Congress*, 6-9 Sep, Santa Rosa, CA.

Contributed oral presentations

- **Dineen, Á.T.,** O'Doherty, C., Truzzi, A., Kravchenko, A., King, G., Wade, A.R., Zaadnoordijk, L., D'Arcy, E.-L., Holloway, T., White, J., Caldinelli, C., Byrne, A.T., Tarrant, A., Foran, A., Molloy, E.J., Cusack, R. (2026). Awake infant fMRI and deep neural network models reveal hierarchical development of object feature tuning in the ventral visual stream. *Cognitive Neuroscience Society (CNS) Annual Meeting*, 7–10 Mar, Vancouver, BC. Data Blitz oral presentation. **Graduate Student Award winner.**

- Ahir, P., **Dineen, Á.T.**, Qin, C., Cusack, R. (2025). Deep spatiotemporal neural networks and tSNR-guided frame retention for frame quality classification in infant fMRI. *BIC-ISMRM Annual Meeting*, 10–11 Sep, Brighton, UK. Oral presentation by Ahir, P.

- O'Doherty, C., **Dineen, Á.T.**, Truzzi, A., King, G., D'Arcy, E.-L., Caldinelli, C., Holloway, T., Molloy, E., Cusack, R. (2025). Features learned from an infant's perspective are more aligned with the infant brain. Fetal, Infant and Toddler Neuroimaging Group Conference, 7–8 Sep, Dublin, Ireland. Oral presentation by O'Doherty, C.

- **Dineen, Á.T.,** O'Doherty, C., Truzzi, A., Kravchenko, A., Wade, A.R., King, G., Zaadnoordijk, L., Holloway, T., White, J., D'Arcy, E., Caldinelli, C., Molloy, E., Foran, A., Cusack, R. (2025). Does the infant ventral visual stream encode global shape or local object features? Trinity College Institute of Neuroscience Early to Late Life Neuroscience Research Day, 4 Jun, Dublin.

- **Dineen, Á.T.**, O'Doherty, C., Truzzi, A., Kravchenko, A., Wade, A.R., King, G., Zaadnoordijk, L., D'Arcy, E.-L., Holloway, T., White, J., Caldinelli, C., Foran, A., Molloy, E., Cusack, R. (2025). Blurry beginnings: the infant ventral visual stream

emphasises low spatial frequency features. *Trinity College Dublin School of Psychology Annual Research Symposium*, 3 Mar, Dublin, Ireland. Poster Spotlight oral presentation.

- O'Doherty, C., **Dineen, Á.T.**, Truzzi, A., King, G., Harrison, K., Zaadnoordijk, L., D'Arcy, E.-L., White, J., Holloway, T., Caldinelli, C., Kravchenko, A., Joseph, S., Molloy, E., Byrne, A.T., Foran, A., Tarrant, A., Cusack, R. (2024). Development of object recognition in the visual cortex from 2 to 9 months using awake fMRI and deep learning. *International Congress of Infant Studies*, 8–11 Jul, Glasgow, UK. Oral presentation by O'Doherty, C.

- O'Doherty, C., **Dineen, Á.T.**, Truzzi, A., King, G., Harrison, K., Zaadnoordijk, L., D'Arcy, E.-L., White, J., Holloway, T., Caldinelli, C., Kravchenko, A., Joseph, S., Molloy, E., Byrne, A.T., Foran, A., Tarrant, A., Cusack, R. (2024). Hierarchical correspondence between deep neural networks and emerging representations in the infant ventral visual stream. *Trinity College Institute of Neuroscience Molecules to Mind Neuroscience Research Day*, 5 Mar, Dublin, Ireland. Oral presentation by O'Doherty, C.

- O'Doherty, C., **Dineen, Á.T.**, Truzzi, A., King, G., Harrison, K., Zaadnoordijk, L., D'Arcy, E.-L., White, J., Holloway, T., Caldinelli, C., Kravchenko, A., Joseph, S., Molloy, E., Byrne, A.T., Foran, A., Tarrant, A., Cusack, R. (2024). Deep neural networks model the development of visual recognition in infants. *Trinity College Dublin School of Psychology Research Symposium*, 7 Mar, Dublin, Ireland. Poster Spotlight oral presentation by O'Doherty, C.

- King, G., **Dineen, Á.T.**, O'Doherty, C., Truzzi, A., Caldinelli, C., D'Arcy, E.-L., White, J., Kravchenko, A., Molloy, E., Foran, A., Tarrant, A., Byrne, A.T., Cusack, R. (2023). Functional connectivity MRI analysis in infancy: is awake richer? *International Neonatology Association Conference*, 7–10 Sep, Dublin, Ireland. Oral presentation by King, G.

- **Dineen, Á.T.**, King, G., O'Doherty, C., Truzzi, A., Caldinelli, C., D'Arcy, E.-L., White, J., Kravchenko, A., Ambre, C., Wadhwa, A., Healy, M., Burke, A., Joseph, S.,

Molloy, E., Foran, A., Tarrant, A., Byrne, A.T., Cusack, R. (2023). Evaluating awake fMRI in one hundred 2-month-olds. *Fetal, Infant and Toddler Neuroimaging Group Conference*, 10–11 Sep, Santa Rosa, CA. **Young investigator award winner.**

- King, G., **Dineen, Á.T.**, O'Doherty, C., Truzzi, A., Caldinelli, C., D'Arcy, E.-L., White, J., Kravchenko, A., Molloy, E., Foran, A., Tarrant, A., Byrne, A.T., Cusack, R. (2023). Infant brain functional MRI: a new awakening. *Irish Neonatal Research Symposium*, 31 Mar, Dublin, Ireland. Oral presentation by King, G.

- King, G., **Dineen, Á.T.**, O'Doherty, C., Truzzi, A., Caldinelli, C., D'Arcy, E.-L., White, J., Kravchenko, A., Molloy, E., Foran, A., Tarrant, A., Byrne, A.T., Cusack, R. (2023). 100 babies: lessons from awake infant fMRI at 2 months. *Trinity College Dublin School of Psychology Research Symposium*, 7 Mar, Dublin, Ireland. Poster Spotlight oral presentation by King, G.

- **Dineen, Á.T.**, Mc Elroy, S., Cusack, R. (2022). Robust motion correction for MRI using DNNs. *Trinity College Institute of Neuroscience Imaging Interest Group*, Trinity College Dublin, 31 Jan, Dublin, Ireland.

- **Dineen, Á.T.**, Cusack, R. (2022). Deep learning for motion robust infant fMRI. *School of Psychology, Trinity College Dublin*, 7 Jun, Dublin, Ireland.

Poster presentations

- Santillo, M.F., O'Doherty, C., **Dineen, Á.T.**, Truzzi, A., Kravchenko, A., King, G., Zaadnoordijk, L., D'Arcy, E.-L., Holloway, T., White, J., Caldinelli, C., Byrne, A.T., Tarrant, A., Foran, A., Molloy, E.J., Cusack, R. (2026). The origins of memorability in infancy. *Cognitive Neuroscience Society Annual Meeting*, 7–10 Mar, Vancouver, BC.

- Conyngham, C., O'Doherty, C., **Dineen, Á.T.**, Truzzi, A., King, G., D'Arcy, E.-L., Caldinelli, C., Holloway, T., Molloy, E., Cusack, R. (2025). The role of structural

connectivity at birth in defining category representations in 2-month-olds. *Fetal, Infant and Toddler Neuroimaging Group Conference*, 7–8 Sep, Dublin, Ireland.

- Hogan, A., O'Doherty, C., **Dineen, Á.T.**, Truzzi, A., King, G., Holloway, T., D'Arcy, E.-L., Caldinelli, C., Molloy, E., Cusack, R. (2025). Can early neural responses to objects in infants' ventral visual stream predict language development? *Fetal, Infant and Toddler Neuroimaging Group Conference*, 7–8 Sep, Dublin, Ireland.

- Santillo, M., O'Doherty, C., **Dineen, Á.T.**, Truzzi, A., Caldinelli, C., D'Arcy, E.-L., Kravchenko, A., Holloway, T., Molloy, E., Foran, A., Tarrant, A., Byrne, A.T., Cusack, R. (2025). Activity evoked in the multiple demand network of two- and nine-month-olds by naturalistic stimuli. *Fetal, Infant and Toddler Neuroimaging Group Conference*, 7–8 Sep, Dublin, Ireland.

- Santillo, M., O'Doherty, C., **Dineen, Á.T.**, Truzzi, A., King, G., Caldinelli, C., D'Arcy, E.-L., Kravchenko, A., Holloway, T., Molloy, E., Foran, A., Tarrant, A., Byrne, A.T., Cusack, R. (2025). Activity evoked in the multiple demand network of two-month-olds by naturalistic stimuli. *Organization for Human Brain Mapping*, 24–28 Jun, Brisbane, Australia.

- **Dineen, Á.T.**, O'Doherty, C., Truzzi, A., Kravchenko, A., Zaadnoordijk, L., Wade, A.R., King, G., Caldinelli, C., D'Arcy, E.-L., White, J., Holloway, T., Molloy, E., Foran, A., Byrne, A.T., Tarrant, A., Cusack, R. (2024). Examining the spatial scale of infant visual representations using deep neural network models and awake fMRI. *Fetal, Infant and Toddler Neuroimaging Group Annual Conference*, 25–26 Sep, Baltimore, MD. **FIT'NG poster award winner.**

- **Dineen, Á.T.**, O'Doherty, C., Truzzi, A., Kravchenko, A., Wade, A.R., King, G., Zaadnoordijk, L., Caldinelli, C., D'Arcy, E.-L., White, J., Holloway, T., Molloy, E., Foran, A., Byrne, A.T., Tarrant, A., Cusack, R. (2024). Probing the effect of blurry vision in infancy on developing object representations in the ventral visual stream using deep neural networks and awake infant fMRI. *Neuro-AI-Talks*, 2–3 Sep, Osnabrück, Germany.

- O'Doherty, C., **Dineen, Á.T.**, Truzzi, A., King, G., Harrison, K., Zaadnoordijk, L., D'Arcy, E.-L., White, J., Holloway, T., Caldinelli, C., Kravchenko, A., Joseph, S., Molloy, E., Byrne, A.T., Foran, A., Tarrant, A., Cusack, R. (2024). The development of object recognition in infancy: findings from neuroimaging and deep learning. *Organization for Human Brain Mapping*, 23–27 Jun, Seoul, South Korea.

- **Dineen, Á.T.**, O'Doherty, C., Truzzi, A., Kravchenko, A., Zaadnoordijk, L., Wade, A.R., King, G., D'Arcy, E.-L., White, J., Holloway, T., Caldinelli, C., Foran, A., Molloy, E.J., Cusack, R. (2024). The effect of low acuity in infancy on developing ventral visual stream representations. *Vision Sciences Society*, 17–22 May, St. Pete Beach, FL.

- O'Doherty, C., **Dineen, Á.T.**, King, G., Truzzi, A., Caldinelli, C., Zaadnoordijk, L., D'Arcy, E.-L., White, J., Kravchenko, A., Ambre, C., Wadhwa, A., Healy, M., Burke, A., Joseph, S., Foran, A., Tarrant, A., Byrne, A.T., Molloy, E., Cusack, R. (2023). Characterising development in the ventral visual stream in young infants with neuroimaging and deep neural networks. *19th NVP Dutch Society for Brain and Cognition Winter Conference*, 14–16 Dec, Egmond aan Zee, Netherlands.

- **Dineen, Á.T.**, King, G., O'Doherty, C., Truzzi, A., Caldinelli, C., D'Arcy, E.-L., White, J., Kravchenko, A., Ambre, C., Wadhwa, A., Healy, M., Burke, A., Joseph, S., Molloy, E., Foran, A., Tarrant, A., Byrne, A.T., Cusack, R. (2023). Evaluating awake fMRI in one hundred 2-month-olds. *Lifespan Neuroimaging Meeting*, 19 Oct, Dublin, Ireland

- O'Doherty, C., **Dineen, Á.T.**, King, G., Truzzi, A., Caldinelli, C., Zaadnoordijk, L., D'Arcy, E.-L., White, J., Kravchenko, A., Ambre, C., Wadhwa, A., Healy, M., Burke, A., Joseph, S., Foran, A., Tarrant, A., Byrne, A.T., Molloy, E., Cusack, R. (2023). Characterising development in the ventral visual stream in young infants with neuroimaging and deep neural networks. *Neuro-AI-Talks*, 24–25 Sep, Osnabrück, Germany.

- **Dineen, Á.T.**, King, G., O'Doherty, C., Truzzi, A., Caldinelli, C., D'Arcy, E.-L., White, J., Kravchenko, A., Zaadnoordijk, L., Burke, A., Healy, M., Joseph, S., Molloy, E.,

Foran, A., Tarrant, A., Byrne, A.T., Cusack, R. (2023). 100 babies: lessons from awake infant fMRI at 2 months. *Organization for Human Brain Mapping*, 22–26 Jul, Montréal, QC.

- **Dineen, Á.T.**, King, G., O'Doherty, C., Truzzi, A., Caldinelli, C., D'Arcy, E.-L., White, J., Kravchenko, A., Ambre, C., Wadhwa, A., Healy, M., Burke, A., Joseph, S., Molloy, E., Foran, A., Tarrant, A., Byrne, A.T., Cusack, R. (2023). 100 babies: an evaluation of awake infant fMRI at 2-months corrected age. *Computational Psychiatry Conference*, 6–8 Jul, Dublin, Ireland.

- **Dineen, Á.T.,** Kravchenko, A., Truzzi, A., Wade, A.R., Cusack, R. (2023). Do blurry beginnings help us see the big picture? Trinity College Dublin School of *Psychology Research Symposium*, 7 Mar, Dublin, Ireland.

List of Abbreviations

AI	Artificial intelligence
AP	Anterior-to-posterior phase encoding
BOLD	Blood-oxygenation-level-dependent
CI	Confidence interval
CGA	Corrected gestational age
CNN	Convolutional neural network
cpd	Cycles per degree
cpp	Cycles per pixel
dva	degrees of visual angle
DNN(s)	Deep neural network(s)
EEG	Electroencephalography
EPI	Echo-planar imaging
EVC	Early visual cortex
FD	Framewise displacement
FG1–4	Fusiform gyrus regions 1–4
fMRI	Functional magnetic resonance imaging
fNIRS	Functional near-infrared spectroscopy
FOUNDCOG	Foundations of Cognition
FOV	Field of view
GLM	General linear model
HSF	High spatial frequency
IQR	Interquartile range
LOC	Lateral occipital complex
LSF	Low spatial frequency
MEG	Magnetoencephalography
MRI	Magnetic resonance imaging
MSF	Middle spatial frequency
NICU	Neonatal intensive care unit
PA	Posterior-to-anterior phase encoding

ROI(s)	Region(s) of interest
ROP	Retinopathy of prematurity
RSA	Representational similarity analysis
RSM	Representational similarity matrix
SF	Spatial frequency
T1w	T1-weighted
T2w	T2-weighted
TE	Echo time
TR	Repetition time
V1	Primary visual cortex
V2–V4	Visual areas 2–4
VEP	Visually evoked potential
VO1	Ventral occipital area 1
VVS	Ventral visual stream

Table of Figures

Figure 2.1 Designing visual paradigms to characterise the ventral visual stream with awake infant fMRI. 65
Figure 2.2 Simulation of how the FOUNDCOG stimuli might appear to a 2-month-old infant given the known perceptual immaturities at this age and the parameters of the display. 71
Figure 2.3 Awake infant fMRI success rates from the FOUNDCOG project 89
Figure 2.4. Reliability of group-level patterns increases as a function of sample size. .. 93
Figure 2.5 Infant attentive state and engagement with stimuli during awake scanning. 95
Figure 3.1 Object feature tuning was characterised along the developing ventral visual stream using awake infant fMRI and deep neural network modelling. 126
Figure 3.2 Development of multiscale object feature tuning along the ventral visual stream. 130
Figure 3.3 Features captured by infant V1 reflect object-specific tuning. 139
Figure 4.1 In untrained deep neural networks, high spatial frequency object features best capture early visual cortex responses across development 183
Figure 4.2 Characterising the band-wise trajectory of convergence with the early visual cortex across development 190
Figure 4.3 Convergence of model representations learned from features at different spatial scales. 193
Figure 4.4 Representations learned from low spatial frequency object features were a relatively better fit for infants, while those learned from high spatial frequency object features were a relatively better fit for adults. 196
Figure A2.1 The amount of awake functional MRI data acquired at the 2-month scan was weakly but significantly predictive of the amount of data that would be acquired at the 9-month scan 261
Figure A3.1 Sample object as seen by an infant viewing the FOUNDCOG experimental display over the first postnatal year. 272

Figure A3.2 Ventral visual stream regions of interest for which tuning to visual object features was characterised. 273
Figure A3.3 Blurrier trained DNNs preferred coarser object features. 274
Figure A3.4 ResNet-50 captures developing object responses along the ventral visual stream. 275
Figure A3.5 Model layers that best capture multiscale object feature tuning along the developing ventral visual stream. 277
Figure A3.6 Multiscale object feature tuning for later visual regions. 279
Figure A3.7 Similarity of multiscale tuning maps along the ventral visual stream across development. 280
Figure A3.8 Correlation between tuning maps and blur-dependent model classification accuracy. 282
Figure A3.9 Multiscale object feature tuning for early visual regions scaled by the noise ceiling. 283
Figure A3.10 Reliability of object feature tuning across the developing ventral visual stream. 284
Figure A3.11 Prematurity was found not to affect the development of multiscale object feature tuning in the early visual cortex at 2 months. 286
Figure A3.12 Replication of multiscale object feature tuning using a low-motion subset of infant runs. 287
Figure A3.13 Replication of multiscale object feature tuning for later visual regions using a low motion subset of infant runs. 289
Figure A3.14 Correspondence between V1 and retinal-wave-trained models, tested on images across a range of blurs. 290
Figure A3.15 Model–brain correspondence for retinal-wave-trained models along the ventral visual stream. 292
Figure A4.1 Preliminary characterisation of the band-wise trajectory of convergence with the early visual cortex across development to estimate approximate point of learning across bands. 298
Figure A4.2 Band-wise trajectory of convergence with the early visual cortex across development. 299

Figure A4.3 Layer-wise trajectories of convergence by band throughout the early visual cortex across development. 300
Figure A4.4 Preliminary characterisation of the band-wise trajectory of convergence with the later ventral visual stream regions across development. 302

Table of Tables

Table 2.1 Definitions of the behavioural tags 87
Table A2.1 Semantic tag definitions 262
Supplementary Table A2.2. Literature-based contrasts of interest 264
Table A3.1 Split-half noise ceilings for regions of interest along the ventral visual stream. 271
Table A4.1 Kruskal–Wallis omnibus tests of spatial-frequency band effects on model–brain correspondence in untrained networks by ROI and age (BH–FDR corrected) 304
Table A4.2 Post-hoc pairwise Mann–Whitney U tests comparing spatial-frequency bands on model–brain correspondence in untrained networks by ROI and age (Holm corrected) 305
Table A4.3 Channel-wise mean and standard deviation of bandpass-filtered ImageNet training images (RGB) 306
Table A4.4 Preliminary band-wise logistic fit parameters across 60 training epochs by ROI and age (1 iteration per band) 307
Table A4.5 Kruskal–Wallis omnibus tests of band effects on preliminary learning dynamics across 60 training epochs by ROI and age (1 iteration per band; BH–FDR corrected) 309
Table A4.6 Post-hoc pairwise Mann–Whitney U tests comparing spatial-frequency bands for preliminary learning dynamics across 60 training epochs by ROI and age (1 iteration per band; Holm corrected) 309
Table A4.7 Band-wise logistic fit parameters early in training (epochs 0–3) by ROI and age (10 model iterations per band) 310
Table A4.8 Kruskal–Wallis omnibus tests of band effects on early learning dynamics (epochs 0–3) by ROI and age (10 model iterations per band; BH–FDR corrected) 312
Table A4.9 Post-hoc pairwise Mann–Whitney U tests comparing spatial-frequency bands for early learning dynamics (epochs 0–3) by ROI and age (10 model iterations per band; Holm corrected) 312

Table A4.10 Within-network paired Wilcoxon signed-rank tests of change in model–brain correspondence from an untrained starting point to the endpoint of the critical training window (epoch 0 vs 3) by ROI, age, and band (10 model iterations per band; BH–FDR corrected across 36 tests) 313
Table A4.11 Kruskal–Wallis omnibus tests of spatial-frequency band effects on model–brain correspondence at the endpoint of the critical training window (epoch 3) by ROI and age (BH–FDR corrected) 314
Table A4.12 Post-hoc pairwise Mann–Whitney U tests comparing spatial-frequency bands on model–brain correspondence at the endpoint of the critical training window (epoch 3) by ROI and age (Holm corrected) 315
Table A4.13 Band-specific differences between infants and adults in model–brain correspondence by ROI (exact sign-flip permutation test; BH–FDR corrected) ... 316

Table of Contents

Declaration ..... **iii**

Summary ..... **v**

Acknowledgments ..... **ix**

List of Publications and Presentations ..... **xii**

Publications ..... *xii*

Manuscripts in preparation ..... *xiii*

Conference proceedings ..... *xiii*

Invited talks ..... *xiv*

Contributed oral presentations ..... *xvi*

Poster presentations ..... *xviii*

List of Abbreviations ..... **xxii**

Table of Figures ..... **xxiv**

Table of Tables ..... **xxvii**

Chapter 1 The Potential of NeuroAI as a Window into the Infant's Visual World **37**

1.1 The infant's visual world **39**

1.1.1 Early visual experience *39*

1.2 The potential of NeuroAI to characterise the visual features encoded by the developing brain **41**

1.2.1 Computational models of developing cognition *42*

1.2.2 Deep neural network models of object processing in the ventral visual stream *45*

1.2.3 The potential of developmentally inspired training regimes to achieve robust, shape-based object recognition *47*

1.2.4 The potential of infant neuroimaging to characterise object responses from the developing brain *49*

1.3 Aims and objectives **50**

Chapter 2 Awake Infant Functional MRI for Visual Neuroscience: Reflections from the Foundations of Cognition Project **53**

2.1 Publication details **57**

2.1.1 Authors and affiliations *57*

2.1.2 Author contributions *58*

2.2 Introduction **59**

2.3 Methods **66**

2.3.1 Design *66*

2.3.1.A Designing and optimising visual paradigms for awake infant fMRI 66

2.3.1.B Considering perceptual development in paradigm design 68

2.3.1.C Identifying perceptual confounds and semantic features in the video stimuli 73

2.3.1.D Shortlisting video stimuli: minimising perceptual confounds and maximising semantic spread 75

2.3.1.E Choosing objects for the pictures task 75

2.3.1.F Final stimulus selection: adult fMRI piloting 76

2.3.2 Implementation *77*

2.3.2.A Participants and ethics 77

2.3.2.B Acquisition of functional MRI 78

2.3.2.C Flexible acquisition protocol 79

2.3.3 Evaluation *81*

2.3.3.A Functional MRI preprocessing and analysis 81

2.3.3.B Reliability of group-level patterns as a function of sample size 82

2.3.3.C Characterising infant attentiveness 85

2.4 Results **88**

2.4.1 Awake fMRI acquisition success rate at 2-months *88*

2.4.2 Awake fMRI acquisition success rate at 9-months *90*

2.4.3 Motion during awake infant fMRI *91*

2.4.3.A Reliability of group-level patterns increases as a function of sample size. 93

2.4.3.B Paradigm effectively captured infant attention 94

2.5 Discussion **99**

Chapter 3 The Development of Tuning to Coarse and Fine Visual Features in Infant Visual Cortex **113**

3.1 Publication details **116**

3.1.1 Authors and affiliations *116*

3.1.2 Author contributions *117*

3.2 Introduction 118

3.3 Results 127

3.3.1 Object responses across adult early visual cortex are best modelled by deep neural networks trained on blurry images 127

3.3.2 Infant V1 is tuned to multiscale object features by 2 months 132

3.3.3 Protracted development of object feature tuning in extrastriate visual cortex 134

3.3.4 What are the possible origins of object tuning in V1? 136

3.4 Discussion 140

3.5 Methods 147

3.5.1 Participants 147

3.5.2 Awake infant fMRI for visual neuroscience 149

3.5.3 Recording object responses with fMRI 150

3.5.3.A Paradigm 150

3.5.4 Deep neural network models 151

3.5.4.A Model training 151

3.5.4.B Characterising the spatial frequency content of the visual features extracted by blur-trained DNNs 152

3.5.5 Representational similarity analysis 152

3.5.5.A Regions of interest 152

3.5.5.B Group-level representational geometry 153

3.5.5.C Characterising visual representations from DNN activations 154

3.5.6 Characterising object feature tuning across spatial scales 155

3.5.6.A Multiscale tuning maps 155

3.5.6.B Developmental change in model–brain correspondence 156

3.5.6.C Associated feature complexity: corresponding DNN layers 156

3.5.6.D Corresponding model accuracy 157

3.5.6.E Scaled covariance between ages 158

3.5.6.F Testing for motion-related effects 158

3.5.7 Effect of prematurity on object feature tuning 158

3.5.8 Correspondence with retinal-wave-trained models 159

Chapter 4 Learning Dynamics of Visual Representations Across Spatial Scales 163

4.1 Introduction 167

4.2 Methods 171

4.2.1 Decomposing images into frequency bands 171

4.2.1.A Implementation details for bandpass filters 171

4.2.1.B Equivalence between the chosen spatial frequency bands and perceptual channels 173

4.2.2 Band-specific deep neural network models 174

4.2.2.A Categorisation accuracy of DNNs trained on different frequency bands 175

4.2.3 Quantifying model–brain correspondence using representational similarity analysis 175

4.2.3.A Participant data 176

4.2.3.B Characterising band-specific visual representations from DNN activations 176

4.2.3.C Calculating model–brain correspondence 176

4.2.4 Model–brain correspondence across spatial-frequency bands in untrained DNNs 177

4.2.5 Learning dynamics of object features across spatial scales 177

4.2.5.A Estimating the approximate trajectory of convergence between model and brain responses 177

4.2.5.B Characterising learning dynamics with a logistic fit 178

4.2.5.C Robustly characterising early learning dynamics 179

4.2.5.D Comparing representational correspondence before and after DNN training 180

4.2.5.E DNN layer comparisons 180

4.2.6 Band-dependent differences between infants and adults 181

4.3 Results 182

4.3.1 In untrained deep neural networks, high spatial frequency visual features best capture early visual cortex responses across development 182

4.3.2 Learning trajectories of object features across spatial scales 184

4.3.2.A Spatial frequency information across spatial scales contains learnable object features 184

4.3.2.B Preliminary band-wise trajectory of convergence with early visual cortex 185

4.3.3 Characterising band-wise learning trajectories 187

4.3.3.A Band-wise learning dynamics 187

4.3.4 Low spatial frequency object features are learned in later layers, while high spatial frequency features peak in early layers 191

4.3.4.A Layer-wise trajectories of correspondence with the brain by band 191

4.3.4.B Convergence of model representations learned from features at different spatial scales 192

4.3.5 Band-dependent differences across development 194

4.3.6 Trends in later ventral visual stream regions 197

4.4 Discussion 197

Chapter 5 Reflections on a NeuroAI Approach to Understanding Infant Visual Development 207

5.1 General discussion 210

5.1.1 Insights into developing visual cortex 210

5.1.1.A Features encoded by developing visual cortex 210

5.1.1.B Investigating learning trajectories and leveraging the full capabilities of NeuroAI 214

5.1.2 Insights for modelling the ventral visual stream with deep neural networks . 216

5.1.2.A Architectural biases 216

5.1.2.B The potential of critical periods 217

5.1.2.C Modelling the trajectory of perceptual development 218

5.1.3 Clinical potential 218

5.1.4 Limitations 219

5.2 A new window into the infant's visual world 221

Bibliography 225

Appendix 2 251

A2.1 Supplementary methods 251

A2.1.1 Calculating the maximum resolvable spatial frequency from display parameters 251

A2.1.2 Simulated time course modelling to minimise perceptual confounds and maximise semantic spread 252

A2.1.2.A Verifying the contribution of perceptual confounds and semantic regressors 252

A2.1.2.B Optimising design efficiency for literature-defined semantic constructs 252

A2.1.2.C Optimising for stimulus distinctiveness 253

A2.1.2.D Greedy backwards elimination strategy 253

A2.1.3 Adult piloting 255

A2.1.3.A Inter-subject correlation analysis 256
A2.1.3.B Final stimulus shortlisting based on adult piloting 257
A2.1.3.C Ethics and data from piloting 257
A2.1.4 Details of paradigms 258
A2.1.5 Video order 258
A2.1.6 Pictures 259
A2.2 Supplementary results 259
A2.2.1 Infants included in analysis of attentive state 259
A2.3 Supplementary figures 261
A2.4 Supplementary Tables 262
Appendix 3 265
A3.1 Supplementary methods 265
A3.1.1 Participant recruitment 265
A3.1.2 Functional MRI setup and acquisition 266
A3.1.3 Functional MRI analysis and preprocessing 267
A3.1.4 Split-half noise ceiling calculation 268
A3.1.5 Deep neural network training implementation details 268
A3.1.6 Characterising the spatial frequency of visual features extracted by blur-trained DNNs 269
A3.2 Supplementary tables 271
A3.3 Supplementary figures 272
Appendix 4 294
A4.1 Supplementary methods 294
A4.1.1 Empirically estimating equivalence between Gaussian blur and Butterworth cutoffs 294
A4.1.2 Impulse image 294
A4.1.3 Candidate filtering and spectral summary 294
A4.1.3.A Estimating equivalent low-pass boundary (σ = 4) 295
A4.1.3.B Estimating equivalent high-pass boundary (σ = 1) 295
A4.2 Supplementary figures 297
A4.3 Supplementary tables 304

Chapter 1

The Potential of NeuroAI as a Window into the Infant's Visual World

1.1 The infant's visual world

1.1.1 Early visual experience

For centuries, scientists, philosophers, and caregivers alike have wondered what infants perceive of the world around them. As adults, we cannot directly experience the world through the eyes of a child, nor are we able to recall what it was once like. Over the past six decades, in an attempt to better understand the visual world of infants, developmental scientists have invented clever techniques to gain insight into the types of visual input that infants are capable of perceiving (Fantz, 1964; Harris, Atkinson and Braddick, 1976; Teller *et al.*, 1986).

As recently as the 1960s, newborn infants were commonly considered to be effectively blind (Braddick and Atkinson, 2011). While this has been disproven, we do know that infants are born into a blurry world, lacking in colour, contrast and fine-detail (Atkinson, Braddick and Braddick, 1974; Harris, Atkinson and Braddick, 1976; Dobson and Teller, 1978; Teller and Bornstein, 1987; Norcia, Tyler and Hamer, 1990). The developmental trajectory of these perceptual capabilities is relatively well understood, with dramatic improvements over infancy generally followed by an extended period of fine-tuning throughout childhood (Leat, Yadav and Irving, 2009; Dekker et al., 2020; Maule, Skelton and Franklin, 2023).

While a deficit view of infancy was once prevalent, it has been proposed that this transient period of perceptual limitation might be adaptive (Turkewitz and Kenny, 1982). Inspired in part by the phenomenon of metamorphosis, whereby a caterpillar is not

viewed as a deficient butterfly but as an integrated organism well adapted for its current ecological niche, Turkewitz and Kenny proposed, by analogy, that infancy might too be "*harmoniously integrated and uniquely adapted.*" They posit that early perceptual limitations may confer downstream advantages, that are not evident in infancy but emerge later in development. This hypothesis has gained traction and has been extended by recent theorists (Vogelsang *et al.*, 2018; Smith *et al.*, 2018; Vogelsang *et al.*, 2024a; Cusack, Ranzato and Charvet, 2024), receiving support from computational modelling (Elman, 1993; Vogelsang *et al.*, 2018).

What infants can see and how this changes as they develop, is important as it constrains what they can learn from their visual environment. The idea that early visual experience may serve as a curriculum for early learning has been extended beyond perception. Recent studies using lightweight head-mounted cameras (Smith *et al.*, 2015; Long *et al.*, 2024) have shown that the content of early visual experience also changes dramatically and systematically with age (Smith *et al.*, 2018), with young infants viewing predominantly faces, whereas older infants see a preponderance of hands (Jayaraman, Fausey and Smith, 2015; Fausey, Jayaraman and Smith, 2016).

Analysis of these scenes recorded from an infant's perspective has shown that, as early as three weeks of age (the youngest age tested), visual input already contains strong structure in basic scene statistics, for example, showing a strong over-representation of cardinal orientations (Petroff *et al.*, 2025). The statistics of these scenes also differ in important ways throughout development, with egocentric scenes recorded from young infants (aged 3–13 weeks-old) dominated by relatively simple, high-contrast edge

structure (Anderson *et al.*, 2024). This is particularly significant as it suggests that infants' well-established preference for high-contrast stimuli in laboratory settings is consistent with the kinds of visual input that they encounter in their everyday life.

While the trajectories of perceptual development have been well characterised, and we are beginning to understand what young infants attend to in their daily lives, a significant gap remains: when viewing real-world stimuli such as objects, what features of these objects does the developing brain encode?

1.2 The potential of NeuroAI to characterise the visual features encoded by the developing brain

The goal of this thesis is to characterise the types of visual features that the developing brain encodes when infants view objects in the world around them. This requires both a measure of brain function–neuroimaging–and a way to interpret the recorded responses–computational modelling. In conjunction, these methods are often referred to as NeuroAI.

Specifically, I propose that awake infant fMRI and computational modelling with deep neural networks are particularly well suited for this task. Awake infant fMRI has rich potential to characterise developing visual responses in the ventral visual stream (VVS) — the pathway that supports object processing in adults (Mishkin, Ungerleider and Macko, 1983; Goodale and Milner, 1992) — while computational modelling with deep

neural networks (DNNs) offers a promising way to probe the resulting responses. Such a NeuroAI approach has significantly advanced our understanding of the adult VVS in recent years. The application of this approach to infants is novel. However, it lies at the intersection of several well-established research programmes.

1.2.1 Computational models of developing cognition

The value of computational models to advance understanding of developing cognition has long been recognised (Simon, 1962; Boden, 1978). Models can complement empirical behavioural research, providing mechanistic insights into information processing and developmental change (Mareschal and Thomas, 2007). Neural network models in particular (Rumelhart, McClelland and Group, 1986), often termed connectionist models, have proven highly prolific in developmental science (Elman, 1996; Quinlan, 2003; Shultz, 2003; Elman, 2005; Mareschal and Thomas, 2007).

Neural networks are a class of computational models that learn to perform tasks by adjusting their parameters through a process termed training. They are composed of many simple units that process in parallel, historically motivated by simplified models of biological neurons (McCulloch and Pitts, 1943; Rosenblatt, 1958). These units are connected by weighted links, forming a network, with the overall pattern of connections referred to as the model's architecture. During training, the model adjusts its weights to improve performance on a task, enabling it to capture structure in the input and to develop internal representations that are learned through experience rather than being fully specified in advance. These models represent information in distributed patterns

of activity across many units, allowing complex behaviour to emerge from interactions between simple components (Hinton, McClelland and Rumelhart, 1986).

Connectionist models have provided insight into the potential mechanisms underlying the development of categorisation abilities in infants (Mareschal, French and Quinn, 2000). Young infants (3-4 months) have been shown to exhibit asymmetry in their categorisation of cats and dogs, such that they dishabituate to a novel dog following familiarisation with cats but show weaker evidence of novelty for a novel cat following familiarisation with dogs (Quinn, Eimas and Rosenkrantz, 1993). A simple connectionist model trained on the same set of exemplars of cats and dogs was found to reproduce this pattern of behaviour, consistent with the idea that this asymmetry could, in principle, arise from general associative learning operating over the statistical structure of the input.

The features used to train this model were manually chosen and measured (e.g. tail length). In a subsequent study, French and colleagues (French *et al.*, 2002) investigated what kinds of visual features might drive this response by passing the exemplars through a V1-inspired filter intended to approximate early visual encoding. In particular, they attenuated high spatial frequencies to mimic limited acuity early in development and used the resulting activation patterns in place of hand-engineered features. This yielded two key observations. Interestingly, filtering was found to reduce variability among exemplars within a category while increasing separation between categories, suggesting that reduced early visual acuity can, in some contexts, support categorisation. Critically, when a similar connectionist model as before was trained on these limited acuity, V1-

based encodings, it similarly reproduced the infant-like categorisation pattern. These results suggest that an early visual encoding could plausibly support the observed behaviour, offering a potential mechanistic account in which early perceptual limitations can shape the emergence of categorisation.

Connectionist models can also provide insight into developmental processes. Building on the theory of Turkewitz and Kenny (1982), Elman (1993) demonstrated that models trained to process complex sentences performed better when they began training with limited working memory capacity. The working memory of these networks was increased with training, following which they outperformed equivalent models that had access to the full 'adult-like' memory capacity from the beginning. This study emphasised the potential adaptive value of processing limits in facilitating learning (Elman, 1993).

Following a similar principle in vision, a model of binocular disparity was found to benefit from pretraining on initially simplified inputs. When the model was initially trained on images that were limited to either coarse (low spatial frequency) or fine (high spatial frequency) features, it outperformed models that had access to the full range of image features from the onset of training. Notably, while the advantage was not specific to the developmentally inspired ordering (commencing with coarse features), both optimal learning strategies benefited from initially simplified input (Dominguez and Jacobs, 2003).

1.2.2 Deep neural network models of object processing in the ventral visual stream

Neural network models have also proven highly valuable for understanding neural responses in the brain. Several innovations have made these models substantially more powerful in recent decades (LeCun, Bengio and Hinton, 2015; Krizhevsky, Sutskever and Hinton, 2012). A class of these models known as deep neural networks (DNNs; LeCun, Bengio and Hinton, 2015; Goodfellow, Bengio and Courville, 2016) has proven particularly effective for modelling neural data (Yamins and DiCarlo, 2016; Richards *et al.*, 2019). In neural networks, processing units are arranged in layers, with models described as deep when they include multiple intermediate ('hidden') layers between input and output. DNNs can have millions of trainable parameters, giving them the capacity to approximate complex functions and learn high-dimensional structure from the statistical properties of large quantities of input data, rather than relying on pre-specified, hand-engineered features (LeCun, Bengio and Hinton, 2015).

In 2014, large-scale DNNs trained to perform object recognition on labelled images were shown to provide effective models of neural response patterns in high-level visual cortex, outperforming earlier classes of models (Yamins *et al.*, 2014; Khaligh-Razavi and Kriegeskorte, 2014). Remarkably, these models were optimised for task performance rather than being explicitly designed to fit neural data (Yamins and DiCarlo, 2016). Many of the models used in these studies had a convolutional neural network (CNN) architecture. CNNs efficiently exploit spatial structure in images by learning filters with shared weights that are applied across spatial locations in the image. Moreover, their hierarchical organisation supports the learning of progressively more complex

representations across layers. Subsequent work has shown correspondence between processing stages in CNNs and the human ventral visual stream (VVS), such that earlier visual cortex is often best captured by earlier layers and later, more object-selective regions by later layers (Güçlü and van Gerven, 2015). While many variants of DNN architectures now exist, CNNs have prevailed as effective models of the VVS (Conwell *et al.*, 2024).

DNNs are now often considered among the most effective models of VVS processing (Storrs *et al.*, 2021) and are at the centre of a thriving research programme, which has been termed *neuroconnectionism* (Doerig *et al.*, 2023). While the scale of these models can decrease their interpretability, it is also their strength, resulting in their remarkable capacity to learn structure from large datasets. They can therefore be thought of as providing a useful vocabulary for interpreting complex patterns in data (Doerig *et al.*, 2023). Indeed, the features encoded by the VVS can also be highly abstract, similarly eluding straightforward interpretation (Ayzenberg and Behrmann, 2022). Neuroconnectionist models have thus been described as occupying a 'Goldilocks Zone' of abstraction, favouring algorithmic clarity at the expense of some biological detail (Doerig *et al.*, 2023). DNNs provide substantial theoretical flexibility, as researchers can systematically vary properties such as the architecture, learning rule, and the statistics of the input data to test theories of information processing (Richards *et al.*, 2019).

A rich toolkit of methods has been developed to enable comparison between the internal representations learned by models and the responses recorded from the brain (Kriegeskorte, Mur and Bandettini, 2008; Storrs *et al.*, 2021; Kaniuth and Hebart, 2022;

Khaligh-Razavi *et al.*, 2017). For example, representational similarity analysis (RSA) constructs an abstract representational space by quantifying pairwise dissimilarities between stimulus-evoked patterns of response. This space can be compared directly between models and brains, without requiring a one-to-one correspondence between individual units across modalities. Due to the methodological challenges associated with acquiring neural data from infants, to date such methods have been sparsely used to probe developing representations.

1.2.3 The potential of developmentally inspired training regimes to achieve robust, shape-based object recognition

Despite their success, deep neural networks (DNNs) also have several limitations as models of human vision (Xu and Vaziri-Pashkam, 2021; Bowers *et al.*, 2022; Linsley and Serre, 2023). These include well-documented behavioural differences between models and humans. For example, DNNs have been shown to be deficient in configural processing (Jacob *et al.*, 2021; Baker and Elder, 2022) and to be biased towards processing local image features as compared to global shape (Geirhos *et al.*, 2018a; Baker *et al.*, 2018; Doerig *et al.*, 2020). They are also typically less robust than the human visual system, exhibiting poor performance on image distortions outside of their training distribution (Geirhos *et al.*, 2018b), and sensitivity to subtle image manipulations that are imperceptible to humans (Nguyen, Yosinski and Clune, 2015).

An exciting line of research proposes that human visual development – particularly early perceptual limitations – could provide valuable insights for building more robust, adult-like vision models, with the reciprocal goal of better understanding development. This

line of work was spearheaded by Vogelsang, Gilad-Gutnick, and colleagues (2018), repopularising the idea that early limitations in development might constitute a feature rather than a bug (Turkewitz and Kenny, 1982; Elman, 1993).

Inspired by the transient period of reduced visual acuity in infancy, training a CNN on blurry images of faces was found to increase effective receptive-field size in the first layer (as estimated from the learned convolutional kernels). This suggests that blurry training encourages models to extract coarser image features from an extended portion of the input image. Pre-training on blurry images of faces was found to produce a model that performed best across a range of blur levels, with slightly reduced performance on sharp images (Vogelsang *et al.*, 2018).

Subsequent work has implemented similar acuity-inspired training regimes (blurry pretraining) for object-recognition tasks. Models exposed to blurry images during training have been reported to rely more on shape for object classification, relative to models trained only on sharp images that also contain fine details (Yoshihara, Fukiage and Nishida, 2023; Avberšek, Zeman and Op de Beeck, 2021; Jang and Tong, 2021). The resulting improvements are generally modest but encouraging.

A current limitation of this approach is that for object-trained models (more so than face-trained models), introducing fine detail later in training has been shown to reduce reliance on coarser features learned earlier, consistent with a continual learning or '*forgetting*' problem (van de Ven, Soures and Kudithipudi, 2024; Jang and Tong, 2021). Recent work has indicated progress towards overcoming this continual learning problem

(Lu *et al.*, 2025b). A mixed training regime that included both blurry and sharp images throughout led to sustained reliance on coarse features (Jang and Tong, 2024), but was less developmentally motivated. Such forgetting constitutes an important point of divergence between model and human learning relative to development and merits further investigation, raising the possibility of further unmodelled factors that might be important for infant learning, such as critical periods of plasticity (Vogelsang, Vogelsang and Sinha, 2026). Beyond acuity, other ways of reducing the effective dimensionality of early visual input, such as diminished colour sensitivity, have been proposed to be similarly advantageous (Vogelsang *et al.*, 2024b).

Significant progress remains to be made to achieve robust, shape-based, human-like vision. As highlighted above however, mechanisms of infant-development may pose particularly useful possibilities for continued investigation. Indeed, very young children have been reported to show more robust object recognition than DNNs (Ayzenberg *et al.*, 2025), despite having experienced orders of magnitude less visual input than is experienced by such models during training (Frank, 2023).

It remains to be tested whether developmentally inspired training regimes yield internal representations that more closely resemble those of the developing brain.

1.2.4 The potential of infant neuroimaging to characterise object responses from the developing brain

In order to compare such models to the developing brain, suitable neuroimaging data is required. Acquiring such data from infants is methodologically challenging as they

cannot follow task instructions, may be fussy or lose attention quickly, and cannot be instructed to remain still (Ellis *et al.*, 2020). To my knowledge, to date, only two published works from outside of our research group have reported comparisons of infant neuroimaging data to the internal representations from DNN models of object recognition using representational similarity analysis (Bayet *et al.*, 2020; Xie *et al.*, 2022). Both studies used electroencephalography (EEG) to record developing responses. This scalp-based method lacks the spatial precision required to resolve localised activity in the ventral visual stream. Thus, to fully leverage a NeuroAI approach to probe developing representations, awake functional MRI data with a rich stimulus set will be required. In Chapter 2, I discuss how acquiring such data from infants is becoming increasingly viable(Deen *et al.*, 2017; Ellis *et al.*, 2020).

1.3 Aims and objectives

In this thesis, I explore the potential of NeuroAI to provide insight into the visual features encoded by the developing brain. A prerequisite to this approach is obtaining suitable measurements of brain function from infants.

Through the Foundations of Cognition (FOUNDCOG) project, we conducted a large, longitudinal neurodevelopment study (n = 134), acquiring awake fMRI from infants, at 2 and 9 months of age, while they viewed visual stimuli. In **Chapter 2**, I share our success rates and report metrics of data quality (motion and attention), evaluating the viability of this method for recording and characterising developing representations at scale.

Leveraging this data, in **Chapter 3**, I investigate the emergence of multi-scale object feature tuning in infant visual cortex. To do this, I compare developing visual representations to representations from deep neural networks (DNNs) that were trained on blurry images. This developmentally-inspired image transformation emphasises global shape while suppressing fine detail. Thus, training DNNs on images that were systematically blurred by varying amounts yields a set of models with graded sensitivity to visual object features, from fine details to coarse shape, enabling me to probe which visual features best explained object responses in the visual cortex across development.

I subsequently sought to further investigate the learning dynamics of visual object features across spatial scale. Thus, in **Chapter 4**, I trained DNNs using a more artificial but controlled image transformation – bandpass filtering – which enabled isolation of object features across spatial scales. By measuring the correspondence between visual cortex responses and DNNs trained on features at each scale throughout model training, the rate of convergence with the brain response could be characterised for visual features across each scale. I examine whether features at different scales converge at different rates and highlight potential inductive biases in DNN models and the developing VVS for learning information at particular scales.

In **Chapter 5**, I reflect on the primary empirical findings from Chapters 2-4, evaluating the success of the awake infant fMRI and DNNs for generating new insights into the features encoded by the developing visual cortex. I conclude by considering the future potential of such NeuroAI approaches for understanding the functional development of the visual cortex and beyond.

Chapter 2

Awake Infant Functional MRI for Visual Neuroscience: Reflections from the Foundations of Cognition Project

A wake infant functional MRI (fMRI) has considerable potential to provide a window into the function of the developing brain. The spatial resolution of fMRI can provide distinct insights into cortical computations that cannot be obtained from behavioural techniques or scalp-based neuroimaging modalities. Acquiring these data, however, is challenging, as infants cannot be asked to pay attention or remain still. Through the Foundations of Cognition Project (FOUNDCOG), we aimed to investigate the feasibility of awake infant fMRI for characterizing developing visual representations at scale. Here, we reflect on how effectively this goal was achieved, reporting acquisition rates, metrics of data quality, and practical insights from the design process that may inform future studies. We developed two passive viewing paradigms to probe the developing ventral visual stream (VVS), the hierarchy of cortical regions that subserves object recognition in adults. These paradigms were designed to balance competing objectives: capturing and maintaining infant attention, eliciting broad VVS responses, and minimising the number of experimental conditions to obtain sufficient statistical power in noisy infant data. The first paradigm comprised six naturalistic videos, chosen to capture attention and encourage stillness. The second comprised 36 object images, which we hypothesised would be less engaging but, if successfully acquired, would enable fine-grained characterisation of VVS representations and comparison with computational models. We acquired awake fMRI from 130 2-month-old and 65 9-month-old infants. Most 2-month-olds (97%) participated in an awake scan, with a median duration of 18.63 minutes. A smaller majority of 9-month-olds (63%) also participated, with a median duration of 13.02 minutes. Although 9-month-olds contributed fewer minutes of data overall, they moved less during scanning, yielding higher retention rates among acquired runs. Both age groups demonstrated interest in the stimuli, although 2-month-olds did so

for a greater proportion of time (median 75.27%) than 9-month-olds (48.52%). Both paradigms elicited reliable stimulus-driven responses across the developing brain, particularly within the VVS. Noise ceiling estimation indicated that larger sample sizes yielded more reliable group-level patterns. Despite its inherent challenges, these findings demonstrate that awake infant fMRI is a viable tool for characterizing developing visual representations at scale.

2.1 Publication details

This chapter is under review for publication in Frontiers in Neuroscience as a contribution to the Research Topic: *'Innovations in neonatal and infant neuroimaging: multi-modal approaches, and developmental insights*.'

2.1.1 Authors and affiliations

Áine T. Dineen[1,2], Clíona O'Doherty[1,2,3], Anna Truzzi[1,2,4], Graham King[1,2], Lorijn Zaadnoordijk[1,2,5], Enna-Louise D'Arcy[1,2], Chiara Caldinelli[1,2], Anna Kravchenko[1,2], Tamrin Holloway[1,2], Jessica White[1,2], Alex R. Wade[6], Aoibhe De Burca[1,2], Maebh Healy[1,2], Amy Burke[1,2], Cameron Murphy[1,2], Katie Herbert[1,2], Claire Ambre[1,7,8], Anisha Wadhwa[1,2], Oscar Markey [1,2], Josué Rico-Picó[1,9], Deirdre O'Hara[1], Angela T. Byrne[10,11], Ailbhe Tarrant[11, 12], Adrienne Foran[13, 14], Eleanor J. Molloy[14-19], Rhodri Cusack[1,2].

[1]Trinity College Institute of Neuroscience, Trinity College Dublin, Dublin, Ireland.
[2]School of Psychology, Trinity College Dublin, Dublin, Ireland.
[3]Department of Psychology, Stanford University, CA, United States of America.
[4]School of Psychology, Queen's University Belfast, Northern Ireland, United Kingdom.
[5] Center for Brain and Cognition, Universitat Pompeu Fabra, Barcelona, Spain.
[6]Department of Psychology, University of York, England, United Kingdom.
[7]Trinity Biomedical Sciences Institute, Trinity College Dublin, Dublin, Ireland.
[8]Columbia University, NY, United States of America.
[9]Mind, Brain and Behaviour Research Centre (CIMCYC), University of Granada, Spain.
[10]Radiology, Coombe Hospital, Dublin, Ireland.
[11]Radiology, Children's Health Ireland Dublin, Ireland.
[12]Radiology, The Rotunda Hospital, Parnell Square, Dublin 1, Ireland.
[13]Neonatology, The Rotunda Hospital, Parnell Square, Dublin 1, Ireland.
[14]Neonatology, Children's Health Ireland, Dublin, Ireland.
[15]Neurodisability, Children's Health Ireland, Dublin, Ireland.
[16]Discipline of Paediatrics, Trinity College Dublin, the University of Dublin, Dublin, Ireland.
[17]Trinity Translational Medicine Institute, St James Hospital
[18]Trinity Research in Childhood Centre, Dublin, Ireland.
[19]Paediatrics, Coombe Hospital, Dublin, Ireland.

2.1.2 Author contributions

Conceptualisation: **Á.T.D.**, C.O'D., A.Tr., G.K., L.Z., O.M., R.C.

Methodology: **Á.T.D.**, C.O'D., A.Tr., G.K., L.Z., E.L.D., C.C., R.C.

Resources: **Á.T.D.**, C.O'D., A.Tr., G.K., E.L.D., C.C., T.H., J.W., A.Ta., A.T.B., A.F., E.J.M., R.C.

Investigation: **Á.T.D.**, C.O'D., A.Tr., G.K., E.L.D., C.C., T.H., J.W., A.R.W., K.H., C.A., A.W., J.R.P., D.O.H., S.J., A.T.B., A.Ta., A.F., E.J.M., R.C.

Project administration: **Á.T.D.**, C.O'D., A.Tr., G.K., L.Z., E.L.D., T.H., J.W., C.M., A.Ta., A.T.B., A.F., E.J.M., R.C.

Data curation: **Á.T.D.**, C.O'D., A.Tr., G.K., A.D.B., M.H., A.B., R.C.

Formal analysis: **Á.T.D.**, C.O'D., K.C., R.C.

Validation: **Á.T.D.**, C.O'D., R.C.

Software: **Á.T.D.**, C.O'D., A.Tr., G.K., R.C.

Writing – original draft: **Á.T.D.**

Writing – review & editing: **Á.T.D.**, C.O'D., A.Tr., G.K., L.Z., E.L.D., C.C., T.H., J.W., A.R.W., A.D.B., M.H., A.B., K.C., C.M., K.H., C.A., A.W., J.R.P., D.O.H., A.T.B., A.Ta., A.F., E.J.M., R.C.

Visualisation: **Á.T.D.**, A.Tr., G.K., A.R.W., R.C.

Supervision: E.J.M., R.C.

Funding acquisition: R.C.

2.2 Introduction

Throughout infancy, the brain develops at an astonishing rate, increasing dramatically in size and grey matter volume (Gilmore, Knickmeyer and Gao, 2018; Bethlehem *et al.*, 2022), and rapidly forming new synaptic connections (Huttenlocher *et al.*, 1982). While their behavioural repertoire is limited (Blumberg and Adolph, 2023), infants are learning a tremendous amount about the world around them (Maurer and Maurer, 1988; Gopnik, Meltzoff and Kuhl, 1999) through a variety of mechanisms, from statistical learning (Saffran and Kirkham, 2018; Smith *et al.*, 2018) to curiosity-driven exploration (Raz and Saxe, 2020). Over this period, infants are forming foundational representations that will later support the rapid emergence of complex behaviours, such as language acquisition in the second postnatal year (Cusack *et al.*, 2018; Cusack, Ranzato and Charvet, 2024; Zettersten, Foushee and Goddu, 2025; Cusack *et al.*, 2025).

Understanding how neural representations emerge over the first year of life requires methods that can both access relevant brain regions and scale to large samples. Given its spatial resolution, in principle, awake infant functional magnetic resonance imaging (fMRI) offers a way to do this. As a model system, we focus in particular on the ventral visual stream (VVS), which facilitates object and scene recognition in adults. Through the Foundations of Cognition Project (FOUNDCOG), we aimed to investigate the feasibility of awake infant fMRI for characterizing developing representations in the VVS across the first postnatal year.

We focused on vision for several reasons, from historical precedent to practical and theoretical advantages. Vision has a long history as a model system in neuroscience. A substantial proportion of human cortex is dedicated to visual function (~20% of cortical surface area; Wandell, Dumoulin and Brewer, 2007), and visual organisation is relatively conserved across species permitting an extensive comparative literature (Kiorpes, 2015). Accordingly, vision has been instrumental in uncovering fundamental principles of cortical function. Most notably, the seminal experiments of Hubel and Wiesel demonstrated how visual experience shapes cortical circuitry and revealed critical periods of plasticity that remain central to current theories of cortical development (Hubel and Wiesel, 1959; Hubel and Wiesel, 1962; Wiesel and Hubel, 1963; Kiorpes, 2015). Vision therefore offers an opportune lens through which to investigate the mechanisms of cortical development and function more broadly.

From a practical perspective, visual stimuli have many strengths. Naturalistic visual stimuli in particular have been shown to capture attention and encourage stillness during scanning in paediatric populations (Vanderwal *et al.*, 2015; Greene *et al.*, 2018) with evidence that these advantages extend to infancy (Ellis *et al.*, 2020; Jessen *et al.*, 2019). Naturalistic stimuli also have the added benefit of supporting multiple concurrent hypotheses, maximising the richness of the responses and the potential information extracted from each scan.

Finally, vision has distinct theoretical strengths. The VVS has been extensively characterised using fMRI in adults and is known to be sensitive to perceptual and semantic visual features along many partially overlapping dimensions (Bracci and Op de Beeck,

2023). A rich toolkit of computational models exists for probing the features driving visual responses, with particular progress in recent years coming from deep neural networks (DNNs; Kriegeskorte, Mur and Bandettini, 2008; Güçlü and van Gerven, 2015; Yamins and DiCarlo, 2016; Bracci and Op de Beeck, 2016). To date, these tools have been sparsely applied to characterise infant visual responses, with a few recent exceptions, using behaviour (Kiat *et al.*, 2022) and electroencephalography (EEG; Bayet *et al.*, 2020; Xie *et al.*, 2022). Further, these models can test hypotheses about infant development pertaining to factors that cannot be experimentally manipulated in vivo (Vogelsang *et al.*, 2018). These models therefore offer an attractive set of tools for probing developing visual responses, provided that suitable fMRI data can be acquired. Through the FOUNDCOG project, we addressed this prerequisite by focusing on the feasibility of acquiring such data at scale.

To contextualize this work, it is useful to consider what has already been learned from behaviour. The past six decades of behavioural research have revolutionised our understanding of infant cognitive abilities across many domains (Aslin, 2007), spearheaded by the pioneering work of Fantz on visual preferences (Fantz, 1961; Fantz, 1964). The looking-time measures that he developed provided the first systematic method for gaining insight into infant competencies. In the domain of vision, these techniques enabled the study of visual function in human infants, overturning the prevailing belief at that time that infants were born blind (Braddick and Atkinson, 2011). Looking-time paradigms have since been used to chart detailed trajectories of perceptual development, such as acuity and contrast sensitivity (Teller *et al.*, 1974; Atkinson, Braddick and Braddick, 1974; Banks and Salapatek, 1976), and have demonstrated that young infants can already

perform complex cognitive behaviours, such as categorisation (Quinn and Eimas, 1996; Mareschal and Quinn, 2001). Despite this remarkable progress, there is a limit to what behaviour can reveal about underlying neural representations. For example, it is unclear whether successful categorisation relies on the same neural substrates in infants as in adults. Indeed, computational models have shown that such behaviours could, in principle, be mediated by patterns of activation in earlier visual regions than are used in adults (French *et al.*, 2002).

Inferring cognitive states from behaviour remains challenging because the developing brain may encode information that does not manifest in behaviour until later. Neuroimaging, therefore, has considerable potential to provide a complementary window into the function of the developing brain (Spelke, 2002; Sirois *et al.*, 2008; Cusack, McCuaig and Linke, 2018; Yates, Ellis and Turk-Browne, 2021), with a distinct set of strengths and challenges accompanying the application of each neuroimaging modality in infants (Turk-Browne and Aslin, 2024). Scalp-based neuroimaging methods, such as EEG and functional near-infrared spectroscopy (fNIRS), are among the most prevalent in their application to infant research (Azhari *et al.*, 2020; Agyeman *et al.*, 2023). These methods have refined our understanding of perceptual development (Norcia, Tyler and Hamer, 1990), revealed neural markers associated with categorisation behaviours (Quinn, Westerlund and Nelson, 2006), and yielded insights into early object processing (Wilcox and Biondi, 2015; Emberson *et al.*, 2017a). However, such scalp-based modalities primarily record neural activity from the cortical surface, limiting access to deeper structures such as the ventral occipitotemporal cortex including regions of the ventral visual stream (VVS). Further, they

do not afford the fine-grained spatial specificity of activation within a region of interest that is obtained with fMRI.

A distinct advantage of fMRI is therefore its spatial resolution, which can provide insights into regional cortical computations that are not attainable with behavioural techniques or surface-based neuroimaging modalities. To date, fMRI has predominantly been conducted while infants are asleep (Mongerson *et al.*, 2017). In this capacity, it has provided insights into the development of functional networks in the infant brain (Fransson *et al.*, 2007; Gao *et al.*, 2017; Fair *et al.*, 2021). Although some paradigms have been successfully adapted for sleeping infants, particularly in the auditory and somatosensory domains (Anderson *et al.*, 2001; Dehaene-Lambertz, Dehaene and Hertz-Pannier, 2002; Wild *et al.*, 2017; Dall'Orso *et al.*, 2018; Nielsen, Graham and Sylvester, 2023), there is a limit to the cognitive capacities that can be tested during sleep, precluding many paradigms. Visual paradigms that probe higher-level representations, for example, require participants to be awake and attending. Until recently, awake infant fMRI was widely considered prohibitively challenging (with one notable exception; Dehaene-Lambertz, Dehaene and Hertz-Pannier, 2002). As for much of developmental psychology, researchers must contend with infant fussiness and inattention, challenges that are exacerbated by the susceptibility of fMRI to motion-induced artefacts (Ellis *et al.*, 2020). Seminal work over the past decade has demonstrated the viability of awake fMRI in human infants (Biagi *et al.*, 2015; Deen *et al.*, 2017; Ellis *et al.*, 2020), unlocking the potential of this technique.

A handful of pioneering studies have used awake fMRI to gain valuable insight into the development of the VVS in infancy. These studies reveal that the coarse structure of the

VVS is in place by at least 5 months of age, as indicated by the presence of retinotopic maps in early VVS regions (V1–V4; Ellis *et al.*, 2021b; Ellis *et al.*, 2025). Awake fMRI has also enabled measurement of category responses to faces, bodies, scenes and toys in cortical regions that will later become strongly category-selective in adults (Deen *et al.*, 2017). Remarkably, infants as young as 2 months already exhibit category selectivity in a subset of these regions (Kosakowski *et al.*, 2022).

Awake infant fMRI studies conducted to date, including those above, along with a handful of others (such as Ellis *et al.*, 2021a; Yates *et al.*, 2022; Yates *et al.*, 2025), have typically used cross-sectional samples with infants spanning broad age ranges, often over several months. Given the demonstrated efficacy of awake scanning, we opted for a comparatively large longitudinal sample (N = 134 infants) with the goal of investigating the feasibility of awake fMRI at scale. To robustly characterise the development of group-level visual representations, we adopted a two-wave longitudinal design with scans at two narrow age bands, at approximately 2 and 9 months corrected gestational age (CGA), corresponding to early and late stages of the first postnatal year.

We designed two passive viewing paradigms, a video task and a pictures task, with the goal of providing insight into the development of VVS representations while balancing three non-orthogonal objectives: (1) eliciting broad VVS responses; (2) limiting the number of conditions to maximise statistical power and (3) maintaining infant engagement. In this work, we highlight the considerations that informed stimulus selection under these competing constraints (Figure 2.1). We subsequently evaluate the data obtained, reporting acquisition rates, data yields, motion distributions, measures of

infant attentive state and the robustness of group-level responses as a function of sample size.

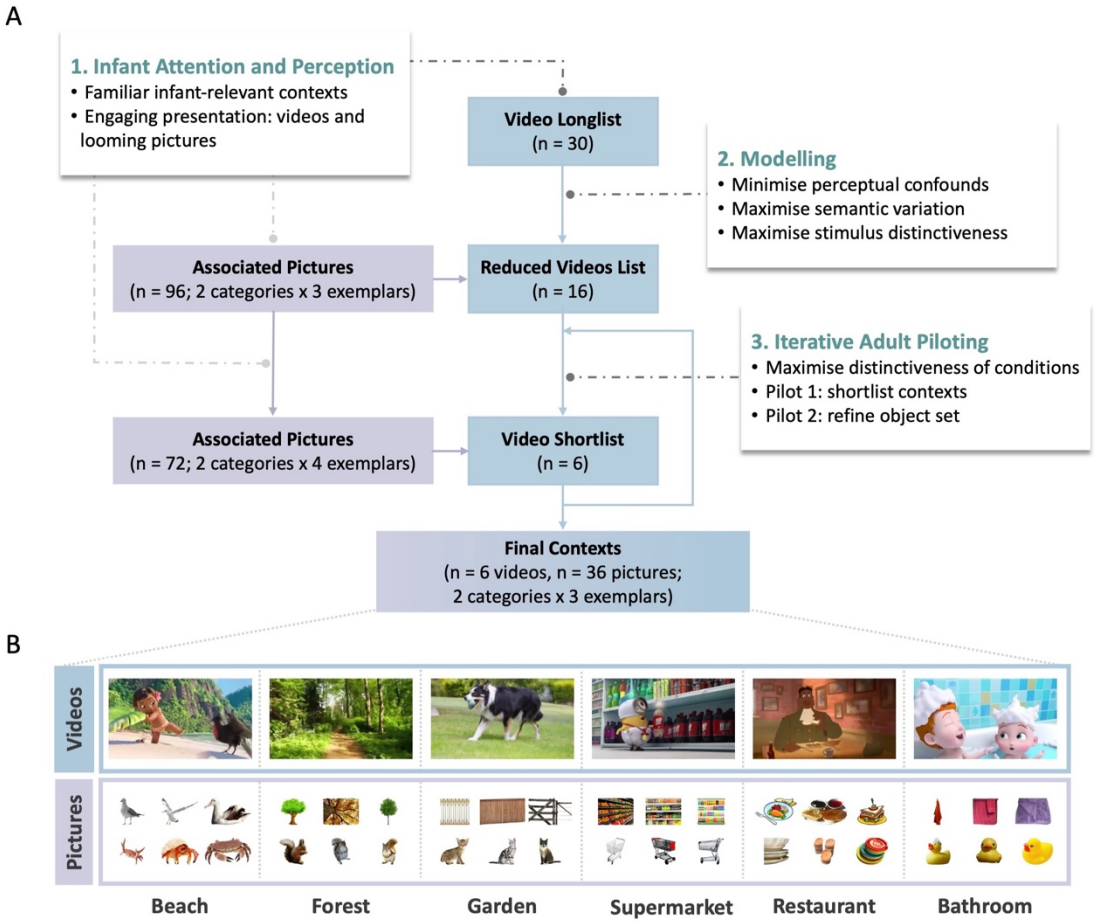

Figure 2.1 Designing visual paradigms to characterise the ventral visual stream with awake infant fMRI.

(A) An overview of the considerations made by the FOUNDCOG team when selecting the optimal stimulus sets for characterising the developing ventral visual stream with awake infant fMRI. (B) The final stimulus sest, optimised to minimise perceptual confounds, maximise semantic coverage, and elicit distinct fMRI response patterns while limiting the number of conditions to maximise power in an infant sample.

2.3 Methods

2.3.1 Design

2.3.1.A Designing and optimising visual paradigms for awake infant fMRI

Our goal was to develop visual paradigms that were sufficiently rich to characterise the developing ventral visual stream (VVS) along semantic dimensions known to drive responses in adults, while also using engaging stimuli that would capture and retain infant attention.

2.3.1.A.a Videos task

Naturalistic visual paradigms have been shown to maximise engagement in paediatric populations (Vanderwal *et al.*, 2015; Richardson *et al.*, 2018). We therefore designed a task comprising a series of short video clips, each depicting a context that an infant could experience over their first postnatal year: in person, through books, or via other forms of media.

2.3.1.A.b Pictures task

We also designed a pictures task. In adults, the VVS is typically investigated using images rather than videos. However, it was anticipated that infants would find images less engaging than videos. An image-based task was therefore considered higher risk, however, if successful, such an acquisition would lend itself to multivariate pattern analysis, enabling computational analyses not previously applied to infant fMRI data. A series of images of objects were therefore chosen. To maximise the utility of the acquisition, related stimuli were chosen for the video and the picture tasks, enhancing the potential for cross-

interrogation. Object categories were therefore selected that were typically associated with the contexts depicted in the videos task and were potentially recognisable by infants.

2.3.1.A.c Initial stimulus selection for the videos task

Infants are a challenging cohort to study as they may become fussy, lose attention quickly, and cannot be instructed to remain still (Ellis *et al.*, 2020). MRI data acquired from infants will therefore inevitably be noisier than that acquired from adults, with a variable number of usable stimulus repetitions across participants. To maximise power, we sought to develop a paradigm with a conservative total number of conditions. We first created an initial longlist of video stimuli (n = 30) that were considered likely to be both visible and engaging to infants. We then assessed whether the content of these videos would be sufficiently visible to infant participants before strategically reducing the number of conditions using simulated fMRI time-course modelling and adult piloting (see sections 2.3.1D–F). An overview of our approach is shown in Figure 2.1A. The videos depicted a variety of contexts, such as mealtime, the playground, and the beach. Redundant contexts were included in the initial longlist to facilitate later optimisation across multiple dimensions.

Videos were edited to standardise their format. Each video was manually cropped to a duration of 22.5 s, which was considered the minimum viable duration required to elicit a reliably measurable and temporally separable hemodynamic response for each individual video regressor. Videos were edited to remove any text by cropping and to standardise framerate (25 frames per second) and aspect ratio (640 × 360 pixels).

2.3.1.B Considering perceptual development in paradigm design

The first postnatal year is a period of rapid perceptual development. Accordingly, consideration was given to the extent to which known limitations of the developing visual system might influence perception of the presented stimuli and thereby confound interpretation of the measured brain responses.

2.3.1.B.a Visual development and stimulus visibility

Visual acuity, contrast sensitivity and colour vision undergo rapid maturation over the first postnatal year (Norcia and Tyler, 1985; Brown and Yamamoto, 1986; Courage and Adams, 1990; Norcia, Tyler and Hamer, 1990; Brown, 1990; Adams, Courage and Mercer, 1991; Teller, 1998; Adams and Courage, 2002). While acuity estimates vary by method, at 4-weeks of age, grating acuity recorded from V1 using visually evoked potentials (VEPs) is typically greater than 5 cpd (Norcia, Tyler and Hamer, 1990).

Stimuli were presented on a rectangular display measuring 29 cm in width and 16.3 cm in height, positioned approximately 27 cm above the infant's eyes. See Methods section 2.2.2 for details of the projection. Although a wider presentation screen was technically feasible, during hardware testing the team judged a larger display to be overly bright and potentially overwhelming. Further, infant peripheral vision is immature. At 2-months, acuity 30 degrees in the periphery is 4–8 times lower than in the centre of vision (Courage and Adams, 1996). Thus, a larger screen would increase the risk that infants might miss salient objects presented in the periphery, which could bias what they attend to, particularly in the video stimuli which can contain many objects on screen at once. It was therefore decided to use a reduced window, resulting in the current dimensions. Given the

fixed aspect ratio of 640 × 360 px, the presented stimuli therefore had a maximum resolvable spatial frequency of 5.7 cpd (see Appendix 2, Supplementary Methods section A2.1.1 for the calculation). Accordingly, by 2 months, infants are likely able to resolve much of the spatial frequency content in our stimuli. Contrast sensitivity, however, is still developing and remains below adult levels, particularly at higher spatial frequencies. By 9-weeks, contrast sensitivity at low spatial frequencies (<1 cpd) reaches approximately 50% of adult levels, while higher spatial frequencies continue to improve throughout development. Thus, at 2 months, contrast for the highest spatial frequencies in our stimuli (1–5.7 cpd) is expected to be attenuated relative to lower frequencies. By 9 months, all spatial frequencies in our stimuli should be resolvable (Hamilton *et al.*, 2021). Therefore, although overall sensitivity continues to increase (Norcia, Tyler and Hamer, 1990), the shape of the contrast-sensitivity function, and therefore the relative perception of the frequencies, is considered adult-like by this age.

Colour vision is also developing throughout infancy. In early infancy, the red–green cone-opponent mechanism becomes functional before the blue–yellow (S-cone opponent) mechanism, which emerges several weeks later (Skelton, Maule and Franklin, 2022). Consequently, for young infants, sensitivity to red and green hues is relatively higher than to blues and yellows, which are harder to detect unless stimuli are large, highly saturated, or are close to the observer (Teller, 1998; Skelton, Maule and Franklin, 2022), such as the ocean in the beach video. At 9 months, although subtle psychophysical immaturities persist, colour vision is largely considered adult-like under everyday viewing conditions (Knoblauch, Vital-Durand and Barbur, 2001; Maule, Skelton and Franklin, 2023).

2.3.1.B.b Simulating infant perception

To visualize the approximate perceptual appearance of our stimuli at 2 months corrected gestational age, we used the TinyEyes simulator (PyTorch implementation), which is openly available at https://github.com/wadelab/VischeckTinyeyes. TinyEyes generates age-based approximations of infant vision, accounting for viewing distance, display size, and age-dependent channel-specific blurring in opponent colour space. Our visual display was relatively large, positioned close to the infant's face, and did not contain high-frequency content (maximum 5.7 cpd). Following simulation with TinyEyes, scene content and object identity remained readily interpretable, suggesting that, given our display parameters, the stimuli were likely sufficiently visible even to the youngest participants (Figure 2.2).

We considered perceptually normalizing the stimuli across participant age groups by transforming all stimuli to match the simulated perception of the youngest participants prior to adult piloting. However, given the limited magnitude of simulated developmental differences and the approximate nature of the modelling process, this was deemed to be unnecessary, and the unmodified stimuli were presented to all age groups.

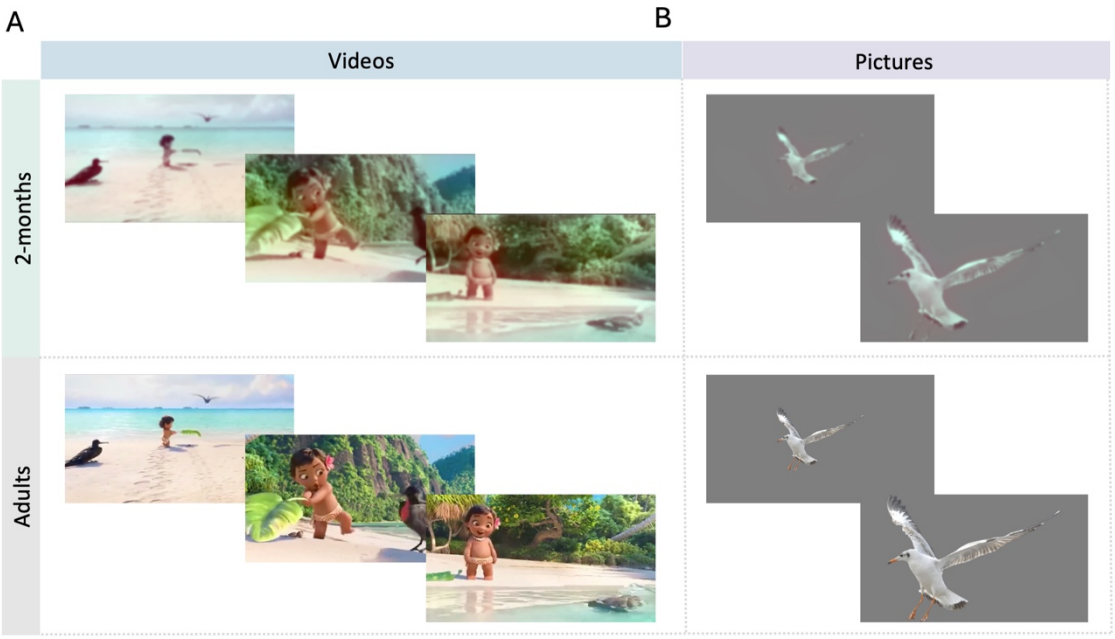

Figure 2.2 Simulation of how the FOUNDCOG stimuli might appear to a 2-month-old infant given the known perceptual immaturities at this age and the parameters of the display.

The TinyEyes simulator was used to approximate the appearance of the stimuli due to known perceptual immaturities in infancy, as a function of the display parameters. Given the large size of the display, the proximity to the viewer, and the relatively low maximum spatial frequency of the stimuli, modelled perception was found to have modest effects on stimulus appearance. The content was deemed to be clearly recognisable at 2-months. (A) Sample frames from the beach context in the videos paradigm, as they might appear to a 2-month-old as compared to an adult. (B) A sample object, seabird, from the pictures paradigm, as it might appear to a 2-month-old as compared to an adult. The seagull is shown at its minimum and maximum loom size relative to the size of the presentation window.

2.3.1.B.c Presenting the objects

Beyond low-level perceptual abilities such as acuity and colour perception, many other aspects of visual development remain immature in infancy. In particular, integrative functions which rely on second-order relations between stimulus features, such as segmentation and grouping, are generally considered to be mediated by visual regions beyond V1 (Braddick and Atkinson, 2011).

In this respect, the pictures paradigm had distinct strengths as it afforded greater control over stimulus presentation. A single segmented object was presented, negating the requirement for the participant to be able to segment the object from its background, a capacity which continues to develop throughout infancy (Braddick and Atkinson, 2011). Objects were presented on a perceptually transparent background. In practice, this was implemented by displaying a black background, as a projector cannot project the absence of light. The object was therefore perceived as appearing directly on the sheet of paper that demarcated the 'display' on the bore of the scanner.

The presentation of a single segmented object meant that, unlike the videos, the pictures were not susceptible to crowding, the inability to recognise an object among visual clutter. Crowding has been shown to have a long time course of development (Farzin, Rivera and Whitney, 2010; Whitney and Levi, 2011). When compiling the initial longlist of videos, contexts which contained only key components with minimal distractor stimuli were therefore preferred. While care was taken to choose videos with minimal visual clutter, in naturalistic videos, particularly those portraying contexts, some degree of clutter is often unavoidable, as such contexts are often defined by the presence of several context-specific

objects. Interestingly, young infants also appear to prefer less cluttered scenes, selectively orienting towards them in their everyday life (Anderson *et al.*, 2024), suggesting that selecting less cluttered scenes might also elicit more attention from young infants.

It was decided that the objects would loom towards the infant. The hypothesised benefits of looming were twofold. Primarily, it was intended to capture infant attention, as was found to be effective for Ellis and colleagues (Ellis *et al.*, 2020). Secondly, behavioural experiments using partly occluded objects have shown that young infants can use motion cues to infer object unity. Thus, looming likely facilitated perceptual grouping of features into a coherent object, increasing the likelihood that young infants would perceive the object as a whole (Braddick and Atkinson, 2011). Objects initially appeared in the centre of the presentation window at half of their maximum height and increased monotonically in size. Size increased according to a logarithmic function such that the stimulus spent the majority of the presentation time near its maximum size.

2.3.1.C Identifying perceptual confounds and semantic features in the video stimuli

2.3.1.C.a Semantic content

To ensure broad coverage of feature dimensions known to drive adult ventral visual stream (VVS) responses, the videos were annotated for object categories, scene categories, and scene characteristics that have been shown to elicit differential responses in the adult VVS. Video clips were manually annotated in ELAN (Sloetjes and Wittenburg, 2008) for the presence of the following categories: faces (Kanwisher, McDermott and Chun, 1997), body parts (Downing *et al.*, 2001), scenes (Epstein and Kanwisher, 1998) and tools (Chao, Haxby

and Martin, 1999); object characteristics: animate, inanimate big and inanimate small (Konkle and Caramazza, 2013); and scene characteristics: near or far, closed or open, nature or civilisation, inside or outside (Kravitz, Peng and Baker, 2011), social or non-social, and biological or non-biological (Huth *et al.*, 2012).

In addition, cinematic characteristics to which the VVS has been shown to be sensitive were annotated, including scene changes and camera cuts (Ben-Yakov and Henson, 2018). Biological motion and salient foreground motion were also tagged, with the latter further coded as moving towards or away from the viewer. Tag definitions are provided in Appendix 2, Supplementary Table A2.1.

2.3.1.C.b Perceptual confounds

To quantify potential perceptual confounds, frame-to-frame root-mean-square (RMS) difference, global contrast factor (Matkovic *et al.*, 2005), and perceptually weighted spatiotemporal spectral energy were calculated for each video. Frame-to-frame RMS difference characterised temporal change in local image statistics by quantifying pixel-wise differences between successive frames. Global contrast factor provided a multiscale summary of spatial contrast by estimating contrast across progressively coarser spatial scales and aggregating these estimates into a single metric per frame. Perceptually weighted spatiotemporal spectral energy characterised spatiotemporal content relevant to human perception by computing the three-dimensional frequency spectrum of each video segment and weighting its frequency components by a previously defined spatiotemporal contrast sensitivity function (CSF) under an assumed viewing geometry.

This CSF was based on adult visual perception and therefore did not capture the developmental differences in perception considered in Methods section 2.3.1.B.

2.3.1.D Shortlisting video stimuli: minimising perceptual confounds and maximising semantic spread

To identify a shortlist of videos that minimised perceptual confounds while maximising semantic spread, several analyses were performed using simulated fMRI time-series modelling. The perceptual and semantic features described above were entered as regressors in a design matrix. This made it possible to identify subsets of videos that minimised perceptual confounds, captured a broad range of semantic dimensions, and maximised discriminability across individual videos. These analyses informed selection of a subset of 16 videos for adult piloting.

2.3.1.E Choosing objects for the pictures task

For each of the initially shortlisted videos (n = 16), two object categories that were typically associated with the depicted context and potentially recognisable by infants were selected. For each object category, three specific exemplars were chosen, yielding a diverse set of 96 unique object images. Where possible, exemplars were selected to vary in viewpoint, to increase the likelihood of capturing view-invariant representations, and to include a mix of naturalistic objects and cartoons.

2.3.1.F Final stimulus selection: adult fMRI piloting

A series of adult pilot studies was conducted to identify the final stimulus set by selecting stimuli that produced the most separable patterns of response in adults. For the pictures task, object images were selected on the basis of both the distinctiveness of the activity patterns they evoked and the similarity of those patterns to the responses elicited by their associated context videos. As it was unclear how long infants would attend to the stimuli in the scanner, the final design was restricted to six contexts, prioritizing a greater number of repetitions over a larger number of conditions.

Two adult pilot fMRI studies were conducted (see Appendix 2, Supplementary Methods section A2.1.3 for further details), which informed selection of the final stimulus set. The final stimulus set comprised six videos, with two object categories per context and three exemplars per object. The final pictures task therefore included 12 object categories, each represented by three unique exemplars. These 12 categories were evenly distributed across animate, inanimate small, and inanimate large classifications, with four object categories in each class.

The final implemented video and picture paradigms are described in detail in Appendix 2, Supplementary Methods section A2.1.4.

2.3.2 Implementation

2.3.2.A Participants and ethics

Ethical approval for the FOUNDCOG project was granted by the Trinity College Dublin School of Psychology Research Ethics Committee (SPREC012020-10, approved 25 February 2020), and the ethics committees of the two collaborating maternity hospitals: the Research Ethics Committee of The Rotunda Hospital (REC-2019-018, approved 31 May 2019) and the Research Ethics Committee in the Coombe Women and Infants University Hospital (Study No. 8 – 2019, approved 8 May 2020). Participation in the FOUNDCOG project included an MRI scan at two timepoints over the first postnatal year, when infants were approximately 2 months and 9 months corrected gestational age (CGA), in addition to behavioural testing and longitudinal behavioural follow-up. Participants were recruited from The Coombe Hospital and the Rotunda Hospital into one of two cohorts, the control group or the neonatal intensive care unit (NICU) group. The recruitment procedure is described in detail in related work (King *et al.*, *in preparation*). Infants in the control group were born at term (37–42 weeks of gestation) and infants in the NICU group were born preterm (born before 37 weeks of gestation) and had spent time in the NICU. For both cohorts, inclusion criteria required the absence of known congenital abnormalities, visual impairments, and contraindications to MRI. Infants in the control cohort were without diagnosed brain injury. Infants in the NICU cohort were excluded if they had retinopathy of prematurity which reached Stage 3 or required laser or injection treatment. For the metrics computed in this work, control and NICU cohorts were combined, giving a single cohort per scan timepoint.

2.3.2.B Acquisition of functional MRI

The hardware setup for the infant scans, as well as the logistics of the scanning session, including research team composition and measures taken to maximise infant and caregiver comfort, are described in detail elsewhere (King *et al.*, *in preparation*).

To present visual stimuli, we used an in-bore projection system inspired by Ellis and colleagues (2020) in which stimuli were projected directly onto the inside of the scanner bore above the infant's eyes. An Optoma UHD30 4K Ultra HD projector, located in the technical room behind the scanner, projected stimuli through an angled waveguide (17° downward from the horizontal) onto an angled front-surface mirror positioned behind the head coil. This mirror reflected the image upwards onto an A4 sheet of paper permanently taped to the bore above the infant's face, which demarcated the presentation window, as in Ellis *et al.* (2020). Prior to presentation, stimuli were geometrically distorted to correct for perceptual distortions introduced by the angled projection and the curvature of the bore. The final projected display appeared as a rectangle measuring 29 cm in width and 16.3 cm in height, located above the isocentre and the infant's eyes. To ensure that infants could see the projected display, scans were acquired using only the posterior array of a Siemens 64-channel head and neck coil (Ellis *et al.*, 2020).

For hearing protection and auditory stimulus delivery, infants wore adult Optoacoustics OptoActive II headphones with the headband removed. The earcups were held in place using inflatable cushions. During awake scans, the headphones were used for passive noise attenuation only, with decibel levels monitored via the system interface. During sleeping scans active noise cancellation was also used.

Active noise cancellation was not used during awake scans because it required a 16 s learning phase during which the headphones sampled the scanner noise. During this phase, auditory stimulus presentation had to be paused, and the infant had to remain still and silent. In practice, this period of silence often led to fussiness and vocalisation, making successful completion of the learning phase difficult. Adult piloting further indicated that under conditions of movement or vocalisation, the system could produce noise-cancellation errors that amplified rather than attenuated scanner noise. As passive attenuation alone provided sufficient hearing protection, active noise cancellation was not used during awake scanning.

Infant fMRI was acquired using a whole-brain T2*-weighted multiband EPI sequence (https://www.cmrr.umn.edu/multiband/). The sequence employed an in-plane acceleration factor of 4, enabling a fast repetition time (TR) of 610 ms. Further acquisition parameters were as follows; echo time of 32 ms, flip angle of 40 0, phase encoding was anterior to posterior, 3 mm isotropic voxels, field of view of 192 × 192 mm, and 64 × 64 × 36 voxels per volume. The video and picture paradigms comprised 485 and 510 TRs respectively, lasting approximately 5 minutes each (4.93 and 5.19 minutes respectively).

2.3.2.C Flexible acquisition protocol

A custom stimulus delivery system was built in PsychoPy (version 2022.1.4, implemented in Python 3.7.12; Peirce *et al.*, 2019), enabling a flexible acquisition protocol. The paradigm is openly available at: https://github.com/ClionaOD/FOUNDCOG_paradigm Before the

infant entered the scanner, a waiting screen comprising dynamic, infant-friendly videos was projected onto the bore. This was intended to capture the infant's attention quickly upon entry into the scanner. The waiting-screen videos were simple and calming, and included colourful fireworks, jellyfish, and abstract patterns similar to those used to encourage compliance in older children (Vanderwal *et al.*, 2015). The research team switched between these videos depending on the infant's apparent preference.

The videos paradigm was always presented first, as it was considered more likely to capture and retain infant attention at the beginning of the session. If a video run was successfully acquired, a pair of single-band spin-echo reference EPI runs, each comprising 10 repetitions and acquired with opposite phase-encoding polarity (anterior-to-posterior and posterior-to-anterior), was then collected for correction of susceptibility-induced distortions. This was followed by the pictures paradigm. If the infant remained engaged, the video and picture tasks were repeated.

If the infant fell asleep, as determined from real-time monitoring of the infant's face using the in-bore camera, the protocol switched to a sleeping sequence consisting of a 10 min resting-state scan followed by T2-weighted and T1-weighted anatomical scans. Further details of the scanning protocol are provided in related work reporting analyses of these datasets (King *et al.*, *in preparation*).

Infants were scanned for a maximum of 40 min at a time, comprising up to 20 minutes of awake scanning and up to 20 min of sleeping scans. They were then removed from the scanner for a break lasting approximately 20 minutes. If the infant and caregiver were

happy to continue, the infant returned to the scanner for a new scan session. In this study, a session was defined as a continuous period during which the infant remained on the scan bed, within the head coil, and with the head in a consistent position inside the scanner (King, Truzzi and Cusack, 2023). Only one visit was conducted per scanning timepoint; thus, when multiple sessions were acquired at a given timepoint, these occurred within the same visit.

When a second session was conducted, infants were again scanned awake for up to 20 minutes. If the infant fell asleep during the second session, the sleeping protocol was followed unless those scans had already been acquired during the first session.

Anatomical scans were not acquired while infants were awake because these three-dimensional sequences were highly vulnerable to motion at any point during acquisition. The sleeping protocol therefore comprised a T2-weighted scan (4 min 5 s) and a T1-weighted 3D MPRAGE scan (6 min 28 s; 3D MPRAGE, FA = 90, TR = 2.3 ms, TE = 2.98 ms, 1 mm-isotropic).

2.3.3 Evaluation

2.3.3.A Functional MRI preprocessing and analysis

Infant fMRI data were preprocessed using a custom Nipype pipeline (Gorgolewski *et al.*, 2011) developed by the FOUNDCOG team for awake infant fMRI. This pipeline is described in detail elsewhere (O'Doherty *et al.*, 2026) and is openly available at https://github.com/ClionaOD/FOUNDCOG_analysis. In brief, standard rigid body motion

correction was performed on the raw EPI volumes using FSL MCFLIRT (Jenkinson *et al.*, 2002) for each functional run. This also provided estimates of framewise displacement (FD) in millimetres between successive volumes. Susceptibility-induced distortions were corrected using FSL topup with field maps acquired in opposite phase-encoding directions (AP/PA; Andersson, Skare and Ashburner, 2003; Smith *et al.*, 2004). Data were then normalised to an age-appropriate template (2-5 month or 8-11 month NIHPD infant template; Fonov *et al.*, 2011) using an affine transformation. Volumes with FD greater than 1.5 mm were excluded, and runs were excluded if more than half of volumes exceeded this threshold.

2.3.3.B Reliability of group-level patterns as a function of sample size

2.3.3.B.a Regions of interest

Regions of interest (ROIs) were defined using cytoarchitectonic parcellations from the Jülich atlas (Amunts *et al.*, 2020a; Amunts *et al.*, 2020b), which have previously been cross-validated against functional regions of the adult ventral visual stream. ROIs were normalised to age-appropriate templates as described in (O'Doherty *et al.*, 2026). Three ROIs were selected for this analysis to sample progressively later stages of the ventral visual stream: V1 (hOc1), hV4 (hOc4v), and VO1 (FG1; Rosenke *et al.*, 2018). For each ROI, voxels from the left and right hemispheres were combined, motivated by previous work showing that homotopic regions exhibit similar visually evoked functional response patterns in young infants (Ellis *et al.*, 2025).

2.3.3.B.b Calculation of representational similarity matrices from the pictures task

Representational similarity analysis (RSA; Kriegeskorte, Mur and Bandettini, 2008) was used to quantify the pairwise similarity structure of ventral visual stream responses to the 36 objects, using Pearson correlation to construct representational similarity matrices (RSMs) and thereby characterise representational geometry.

For each pictures run, a volumetric whole-brain general linear model (GLM) was fit in Nilearn (version 0.9.2; Contributors, 2022). The design matrix included one regressor for each of the 36 objects, convolved with a canonical hemodynamic response function. Nuisance regressors comprised the six motion parameters, a linear trend regressor, and a cosine drift model with a high-pass filter of 0.01 Hz. Volumes with FD greater than 1.5 mm were censored from analysis (Power *et al.*, 2014). Runs were excluded if more than half of volumes exceeded this threshold or if the researcher attending the scan judged the run to be of insufficient quality based on observation.

For each ROI, voxel-wise parameter estimates (betas) were calculated for each object within each run, yielding multivariate response patterns for RSA. Voxels with unstable parameter estimates were excluded. Within each age group and ROI, pairwise Pearson correlations were computed between multivariate object-response patterns for each unique subject-run pair, yielding RSMs. The final sample comprised 101 2-month-old infants and 44 9-month-old infants.

2.3.3.B.c Calculation of temporal representational similarity matrices from the videos task

For each videos run, voxel timeseries were extracted within each ROI. These timeseries were high-pass filtered (0.01 Hz), mean-centred, and scrubbed by setting high-motion timepoints (FD > 1.5 mm) to missing values. Timeseries were then segmented into trials corresponding to individual video presentations (22.5 s each), shifted by 4.5 s to account for hemodynamic delay. Trials with more than 50% scrubbed timepoints were excluded.

Remaining trials were resampled via cubic interpolation to a canonical temporal grid (37 timepoints at 0.610 s resolution), then aggregated across repetitions within each participant using the median while ignoring missing values. Timepoints were averaged within 3.75 s bins (approximately 6 TRs), yielding six temporal bins per video. Bins from all six videos were then concatenated to form a 36-bin temporal pattern for each participant. Patterns were normalised (row-wise demeaning and L2 normalisation) such that cross-participant dot products corresponded to Pearson correlations. This produced a 36 × 36 temporal RSM for each unique participant pair. The final sample comprised 116 2-month-old infants and 47 9-month-old infants.

2.3.3.B.d Bootstrap resampling procedure

To estimate the reliability of group-level functional activation patterns as a function of increasing sample size (n), the noise ceiling was calculated for groups of n participants using a bootstrap resampling procedure for three ventral visual stream ROIs: V1, hV4, and VO1.

For each task, ROI, and sample size n, participants were sampled with replacement across 1000 bootstrap iterations, taking the relevant subject-run pairs from the pairwise RSMs (pictures task) or pairwise temporal RSMs (videos task), as appropriate. For each bootstrap sample, split-half reliability was calculated by dividing the sample into two subgroups, computing the RSM for each subgroup, and correlating the two RSMs. To correct for the reduction in sample size introduced by splitting the data, the Spearman–Brown prophecy formula was applied.

2.3.3.C Characterising infant attentiveness

During the scan, an MRC Systems High Resolution MRI-compatible camera was located above the infant's head to the side of the bore, capturing their facial expression. This enabled real-time monitoring of the infant from the control room. During awake scans, this footage was recorded, enabling retrospective tagging of behavioural state. The camera had a resolution of 1280 by 960 pixels at a framerate of 43 Hz. Initially, videos were manually recorded and cropped to the nearest second to facilitate approximate alignment with the fMRI acquisition. Later in the study, the capacity for automatically starting the recording along with the fMRI acquisition was added to the stimulus delivery system, enabling synchronisation with millisecond precision.

This footage from each recorded run was tagged for several behaviours, including interest, fussiness and looking away, as defined in Table 2.1. For runs where recordings were available, the proportion of time that each tag was applied to each run was calculated to the nearest second. The distribution of the proportion of time that infants performed a

behaviour during individual runs was then compared across age groups (Figure 2.5B), and across task within an age group (Figure 2.5C-D). Runs were included irrespective of the length of the run or the duration of the recording so as not to bias the distributions towards infants who tolerated scanning well. As each infant contributed a variable number of runs, to compare the run-wise distribution of tags across age groups, or across task within an age group, an ordinary least squares regression was used with subject-level clustering to adjust standard errors for within-subject correlations. As multiple behaviours were assessed, p-values were corrected for multiple comparisons using the Benjamini-Hochberg False Discovery Rate method. A further analysis was conducted comparing a subset of participants who had all completed a particular subset of runs (Figure 2.5E-F). In this case, as each infant contributed exactly one run for each condition, within-subject differences in behavioural distributions were assessed using the non-parametric Friedman test. Where significant effects were detected, pairwise Wilcoxon signed-rank tests were used for post-hoc comparison.

Table 2.1 Definitions of the behavioural tags

Behavioural tag	Definition of tag
Interest (interested in and eye scanning stimulus)	A continuous time period when an infant is conducting eye scanning of the visual stimuli. This is not just when the eyes/pupils are physically moving but includes the time between eye movements when the infant appears to be in a continuous state of interest, with intermittent eye scanning movements. This is not simply when the infant is looking at the stimuli.
Fussy	When the infant appears fussy. This tag is given if at least 2 of the following are present in the facial expression: Brow bulging/lowering: bulging, creasing, above and between brows resulting from lowering and drawing the brows together. Nasolabial furrowing: pulling upwards and deepening in the nasolabial fold. Opening lips/mouth: any separation of the lips (even when an object is present in the mouth). Eye squeeze (eyelid must close): squeezing/bulging of the eyelid.
Sleepy	When an infant appears quiet, as if heading towards sleep. A common sign is when eyelids begin to close. However, their eyelids must still be partially open (not fully closed). The infant must not be '*asleep*'
Asleep	When an infant has their eyes fully closed for more than 1 s, and their lack of movement suggests they are asleep. Note: a sucking reflex can occur in infants when asleep.
Eyes closed awake	When an infant has their eyes fully closed for more than 1 s. Sleepy/Asleep preclude the use of this tag.
Looking away	The infant's eyes are deviated such that it appears as though they are not looking towards the stimuli.
Eyes blocked	When the infant's eyes are not sufficiently visible to allow confident tagging. This may be due to an obstructed or unclear camera view, for example if an object is blocking the lens.

2.4 Results

2.4.1Awake fMRI acquisition success rate at 2-months

A total of 134 infants were recruited into the FOUNDCOG Project (54 female; median CGA = 2.3, IQR = [2.13, 2.62]; Figure 2.3B) and attended the 2-month scanning session. Awake fMRI was successfully acquired from 130 2-month-olds (97%), greatly surpassing the expectations of the research team. No data was acquired from three 2-month-olds. One infant successfully participated in an awake acquisition, however, due to a technical error, the acquired data was not usable. Of the infants from whom awake fMRI was successfully acquired, the median duration of awake task-based scanning was 18.63 minutes. The videos task was always conducted first. 113 (87%) infants progressed to the pictures task. Infants participated for longer in the videos task (median 9.86 minutes or 2 full runs), than the pictures task (median 7.97 minutes or 1.5 runs). However, this was at least partially due to ordering effects and therefore cannot be directly interpreted as an expression of stimulus preference. Surprisingly, 104 2-month-olds (80%) also fell asleep enabling the acquisition of a resting state scan (completed by n = 100) and a T2-weighted anatomical scan (103 infants). Later in the study, a T1-weighted scan was also added to the protocol and was therefore acquired from a smaller number of infants (n = 38).

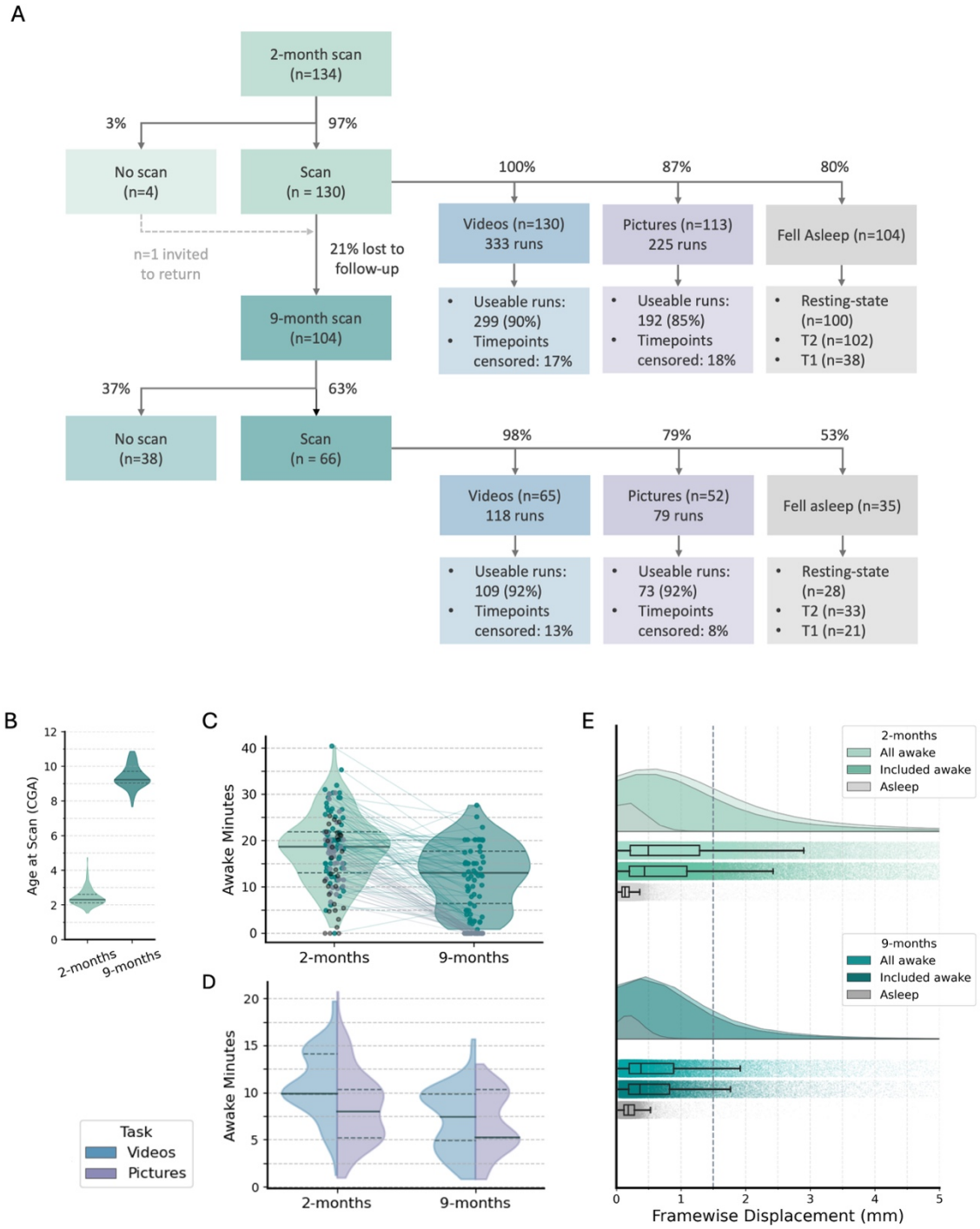

Figure 2.3 Awake infant fMRI success rates from the FOUNDCOG project

The FOUNDCOG project included MRI scanning sessions at approximately 2 and 9 months. (A) Flowchart summarising scan participation and data yield at both timepoints. For infants who completed an awake scan, the number of runs acquired for each visual paradigm (videos and pictures) is shown, together with the proportion of usable runs retained after motion thresholding and the proportion of

timepoints censored within those retained runs. The number of infants who fell asleep during scanning and subsequently completed resting-state fMRI and anatomical scans is also shown. At 2 months, 130 infants completed an awake scan and 104 returned for the 9-month session. At 9 months, 66 infants contributed MRI data, of whom 65 completed an awake scan and 1 contributed sleep-only data. (B) Distribution of corrected gestational age at scan for participating infants at each timepoint. (C) Distribution of total minutes of awake fMRI acquired for each infant at 2 and 9 months, with lines linking infants who contributed data at both timepoints. Teal dots indicate infants who returned at 9 months and contributed awake fMRI data; grey dots indicate infants who returned but did not contribute awake fMRI data; black dots indicate infants who participated at 2 months but were lost to follow-up. (D) Distribution of minutes of awake fMRI acquired for the videos and pictures paradigms at each timepoint. As the videos task was always conducted first, longer participation in the videos task should not be interpreted as an indicator of stimulus preference. (E) Raincloud plots showing framewise displacement (FD) for all awake timepoints, awake timepoints retained for analysis, and resting-state scans acquired during natural sleep, separately for the 2- and 9-month sessions. The dashed vertical line marks the chosen motion threshold of 1.5 mm. In retained awake runs, timepoints with FD > 1.5 mm were censored from analysis. The tails of the awake distributions extended beyond the plotted range but were clipped at 5 mm for visualisation.

2.4.2 Awake fMRI acquisition success rate at 9-months

Infants were invited to return for a 9-month scan if any MRI data was acquired at 2-months. 104 of the 131 (79%) returned (44 female, median CGA = 9.23 months, IQR = [9.03, 9.71], Figure 2.3B) with 27 (21%) lost to follow-up. As expected, success rates at this older timepoint were lower, with awake data successfully collected from 65 of the 104 (62.5%) 9-month-olds (30 female, CGA=9.2, IQR = [9.0, 9.6]). At this timepoint, infants who chose to participate contributed a median of 13.02 minutes of awake data (Figure 2.3C),

comprising a median of 7.42 minutes of videos (1.5 runs) and a median of 5.18 minutes of pictures (1 run; Figure 2.3D). Fewer infants fell asleep at 9-months, with 34 falling asleep during an awake scan (52%) and one infant who did not tolerate awake scanning but fell asleep outside of the scanner and subsequently completed the sleeping scans. Resting-state (completed by n = 28), T1-weighted (n = 21), and T2-weighted (n = 32) scans were acquired from a small number of participants at this age. The number of minutes of awake data at 2 months was found to be weakly but significantly predictive of the number of minutes acquired at 9 months (Spearman's r = 0.22, p = 0.0275), explaining approximately 5% of variance (Appendix 2, Supplementary Figure A2.1).

2.4.3 Motion during awake infant fMRI

While 9-month-olds were less likely to enter the scanner and participated for a shorter duration on average, when they did participate, they were found to move less than 2-month-olds (Figure 2.3E). Across all acquired awake timepoints, 2-month-olds had a median framewise displacement (FD) of 0.55 mm, IQR = [0.25, 1.12], moving slightly less during the videos (median of 0.53 mm, IQR = [0.25, 1.01]) than the pictures (median of 0.56 mm, IQR = [0.25, 1.26]). A motion threshold of 1.5 mm was selected to balance data retention with motion scrubbing, in line with the rationale of Ellis et al., (2020), see Discussion. Runs with a median FD greater than this threshold were excluded from analysis. At 2-months, following thresholding, 299 (90%) video runs (median FD of 0.44 mm, IQR = [0.24, 0.83]) and 192 (85%) picture runs remained (0.45 mm, IQR = [0.24, 1.01]). In the remaining runs, timepoints with a FD of greater than 1.5 mm were censored from analysis (Power *et al.*, 2014), consisting of 17% of timepoints in the videos and 18% in the pictures runs at 2 months.

Across all acquired awake timepoints, 9-month-olds had a median framewise displacement (FD) of 0.42 mm, IQR = [0.25, 0.98], moving less during the pictures (median of 0.32 mm, IQR = [0.22, 0.79]) than the videos (median of 0.55 mm, IQR = [0.28, 1.03]). Following thresholding 109 (92%) video runs (0.49 mm, IQR = [0.27, 0.92]) and 79 (92%) picture runs remained (0.30 mm, IQR = [0.20, 0.57]). Thirteen percent and 8% of timepoints were censored from the remaining videos and pictures runs respectively.

As anticipated, infants moved substantially less when they were asleep, with a median FD of 0.13 mm, IQR = [0.09, 0.20] at 2-months, and 0.18 mm (IQR = [0.12, 0.28]) at 9-months, however, the motion distributions during awake scanning were deemed to be acceptable.

Beyond FD-based exclusions, a further proportion of runs, up to 15% depending on the analysis, were manually excluded due to run-level quality issues, including early termination, infants falling asleep, or other concerns noted manually during the scanning session (O'Doherty *et al.*, 2026).

2.4.3.A Reliability of group-level patterns increases as a function of sample size

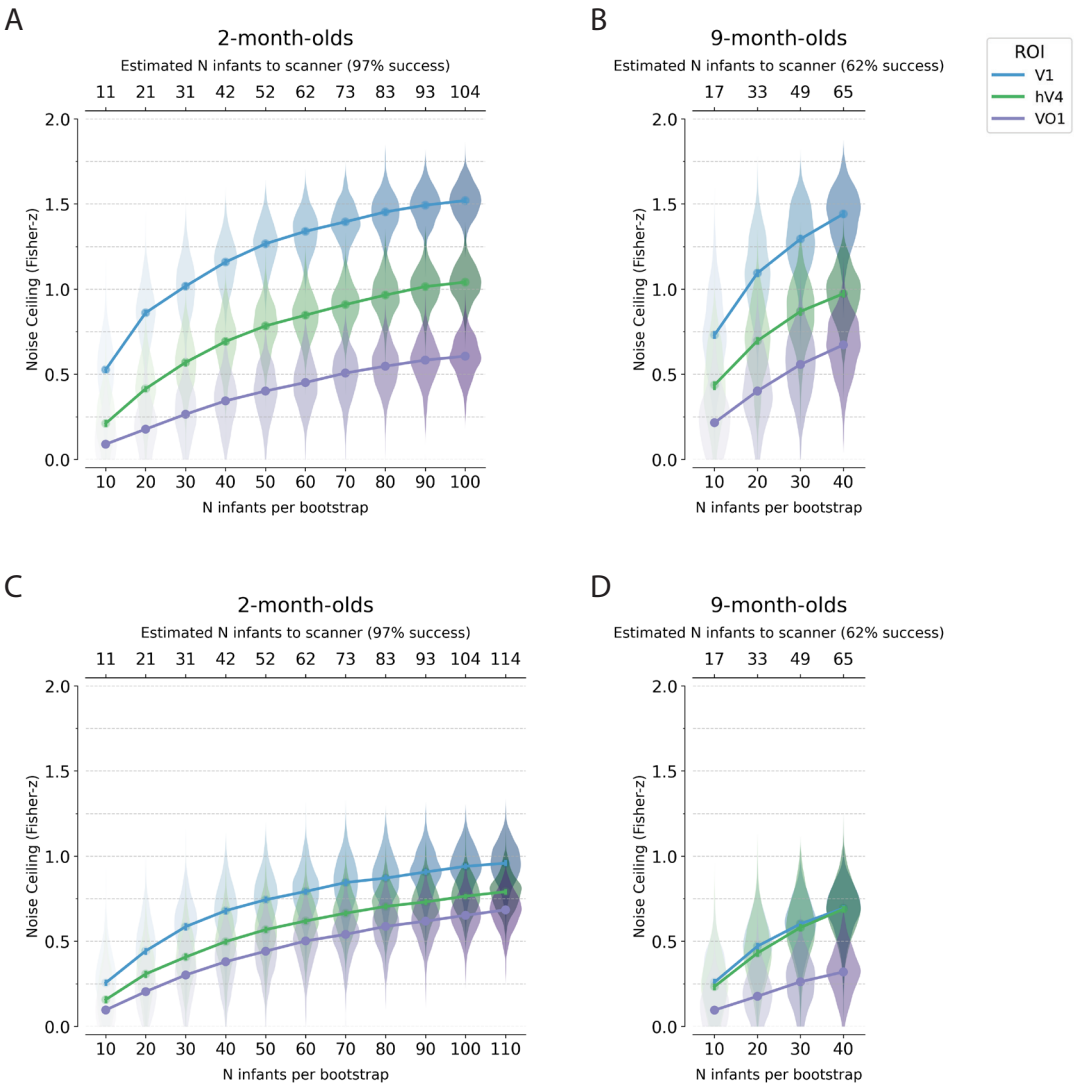

Figure 2.4. Reliability of group-level patterns increases as a function of sample size.

The reliability of group-level patterns was estimated by calculating the noise ceiling for bootstrapped subsets of infants at varying sample sizes in three ROIs along the ventral visual stream: V1, hV4, and VO1. For the pictures task (A, B) and the videos task (C, D), the noise ceiling was estimated by calculating split-half reliability within each bootstrapped sample, then correcting for the reduction in sample size introduced by the split using the Spearman–Brown prophecy formula. Resulting correlation values were Fisher z-transformed. Reliability increased with sample size

for both tasks. Violin plots show the distribution of noise ceiling estimates across 1,000 bootstrapped samples, and the line plot connects the median of each distribution. In each panel, the lower x-axis shows bootstrapped sample size, while the upper x-axis estimates the number of scanning sessions required to obtain a sample of that size for infants in that age group.

A particular strength of the FOUNDCOG project was the large sample size acquired for a given awake infant fMRI paradigm. To assess whether larger samples conferred benefits for analysis, we estimated the noise ceiling (split-half reliability) at increasing sample sizes using a bootstrap resampling procedure in a subset of VVS ROIs (V1, hV4, VO1). Increasing sample size was associated with improved split-half reliability for both the pictures and videos tasks (Figure 2.4). Across all ROIs examined, reliability appeared to increase more gradually across the sampled range for the pictures task, whereas gains for the videos task were more modest at larger sample sizes.

2.4.3.B Paradigm effectively captured infant attention

In-scan facial recordings were available for tagging of attentive state from 93 2-month-olds (369 runs; 233 video runs, 136 picture runs) and 46 9-month-olds (138 runs; 80 video runs, 58 picture runs) totalling approximately 30 hours of awake scan time at 2-months, and 11 hours at 9 months (see Appendix 2, Supplementary Results section A2.2.1 for a breakdown of included participants).

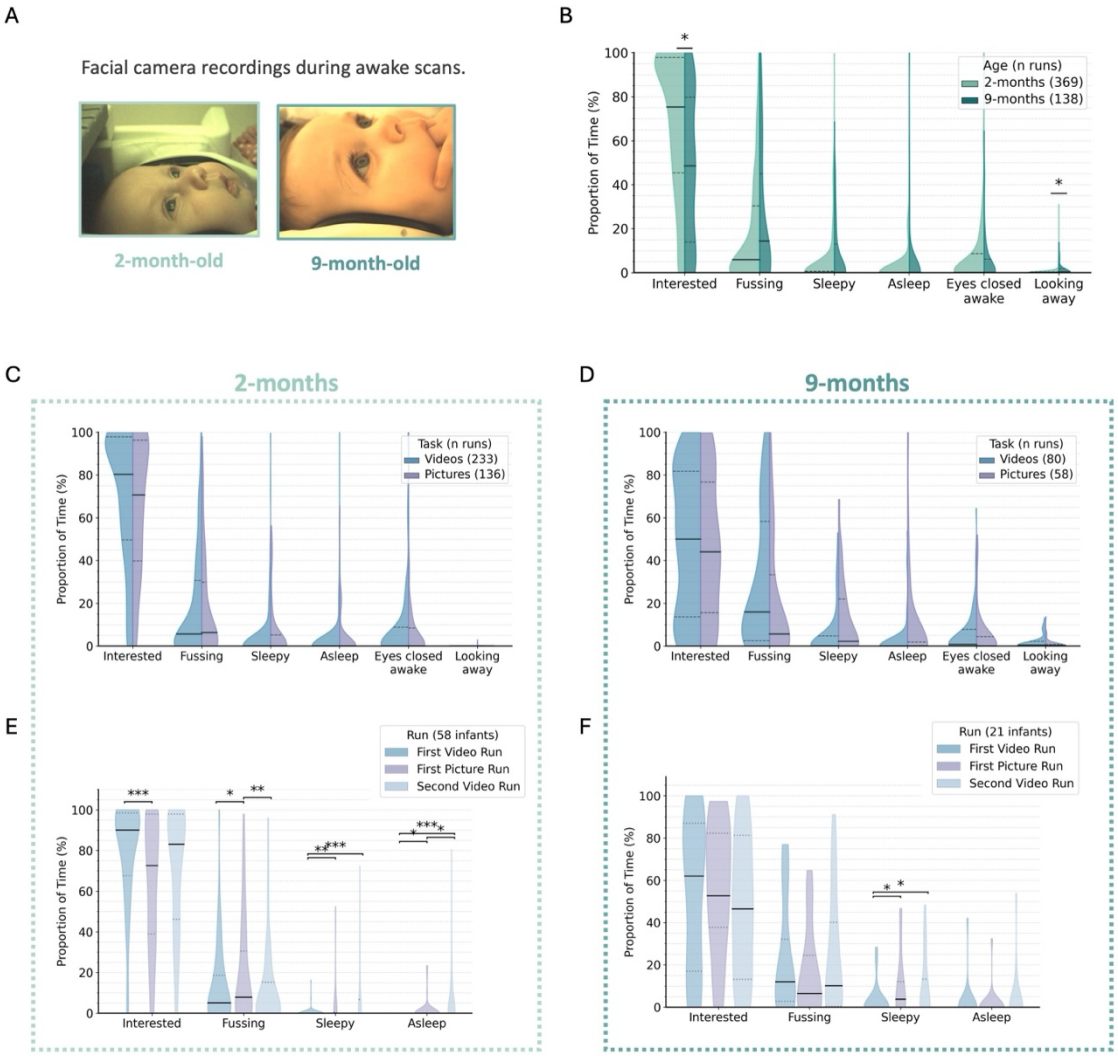

Figure 2.5 Infant attentive state and engagement with stimuli during awake scanning.

During awake fMRI scans, infants' facial expressions were recorded, enabling retrospective tagging of attentive state. The distribution of the proportion of time that an infant performed a behaviour during a given run is depicted by the violin plots. The thick black line through each violin depicts the median proportion across runs, while the dashed black lines depict the inter-quartile-range. Where violins were statistically significantly different from each other, this is depicted by a thick black line above the relevant pair of violins, with the magnitude of significance indicated by the number of asterisks above the line. (A) Sample images from the in bore facial camera during scanning. Images depict a 2-month-old infant and a 9-month-old infant attending to the FOUNDCOG stimuli. Parental

consent was provided to include recognisable images of these infants in scientific publications. (B) Comparison of attentive state by age. 2-month-olds were found to attend to the stimuli for a significantly greater proportion of the time than 9-month-olds. (C-D) The distribution of the proportion of time that an infant spent performing a given behaviour during a run of a given task at 2 months (C) and 9 months (D). (E, F) For infants who participated in the first video run, the first picture run and the second video run, the distribution of the proportion of time that an infant spent performing a subset of behaviours for each of these run-types is depicted for 2-month-olds (E) and 9-month-olds (F).

Across runs, 2-month-olds tended to express interest in the visual stimuli for the majority of scan time (median proportion 75.27%, IQR = [45.42-97.87]). 9-month-olds were also found to be interested in the stimuli (median proportion 48.52%, IQR = [13.95-79.72]) but significantly less so than at 2 months (corrected p-value < 0.00005). 9-month-olds looked away from the presented stimulus for a greater proportion of the time than 2-month-olds. However, the total amount of time looking away remained small (median 0.38% at 9-months, 0% at 2-months; corrected p-value 0.002). There was a trend for 9-month-olds to be fussier (median 14.35%, IQR = [0.08-44.93]) than 2-month-olds (median 5.88%, IQR = [0.0-30.39]) but this difference did not reach significance (uncorrected p-value = 0.056, corrected p-value = 0.113). The proportion of time spent in a state of sleepiness, asleep, or with eyes closed while awake was minimal for most infants, with no significant difference between ages (Figure 2.5B). Note that interest was not simply defined as looking towards the stimulus but rather referred to a continuous period of time where the infant was conducting eye movements, indicating that they were scanning the visual stimulus. Indeed, the proportion of time spent looking away (2 months: median 0.0%, IQR =[0.0-0.35]; 9 months: median 0.38%, IQR = [0.0-1.73]), or with eyes closed while awake

(2 months: median 0.0%, IQR = [0.0-8.64]: 9 months: median 0.0%, IQR =[0.0-6.18]) were found to be minimal for both age groups. The proportion of time spent with eyes closed while awake was slightly higher for a subset of infants but note that this tag was not mutually exclusive with fussing.

Comparing runs by task revealed that 2-month-olds appeared to be more interested in the videos task (median 80.28%, IQR = [49.65-97.89]) than the pictures task (70.66%, IQR = [39.82-96.23]). However, this difference did not reach significance. Indeed, none of the tagged behaviours exhibited significant differences across tasks at 2-months (Figure 2.5C). As the videos were always first, the effect of the stimulus was confounded by the presentation order. From this, it is therefore not possible to dissociate effects of cumulative scan time from those of stimulus-driven attention. To more directly investigate the infants' response to a given task, tasks were therefore also compared in a subset of infants who participated the first videos run, the first pictures run, and the second videos run (n = 58 2-month-olds, 62% of sample contributing to the analysis of attentive state; Figure 2.5E). This enabled comparison of the first video and the first picture task, while excluding infants who fussed out of the session, or fell into a sustained sleep and therefore switched to sleeping scans, during either of these runs. By analysing this subset of tasks, we could somewhat disentangle the response to the stimulus from sustained compliance with the scan. It was found that 2-month-olds were significantly more interested in the first video run than the first picture run ($p = 0.0002$). Infants also fussed less during the first video run (p value = 0.042) than the first picture run. As these infants all progressed to a second video run, they were not sufficiently fussy during the pictures as to end the session. Further, fussing significantly decreased ($p = 0.008$) when returning to the videos

task for the second time. The median interest was also found to increase. However, the distribution was not significantly different from the preceding picture run (p = 0.463). These results indicate that the 2-month-olds significantly preferred the videos task over the pictures task.

At 9-months, across all runs, infants did not appear to be more interested in one task than another (median for videos 50.05%, IQR = [13.62-81.71]; median for pictures 44.05%, IQR = [15.59-76.71]; Figure 2.5D). However, 9-month-olds were fussier during the videos (15.94%, IQR = [2.52-58.28]) than during the pictures (5.64%, IQR = [0.0-33.36]). This difference was found to be significant but did not survive correction for multiple comparisons (uncorrected p-value 0.045, corrected p-value 0.091). 9-month-olds were found to be sleepy, and to fall asleep, significantly more during the pictures than the videos. These differences did not survive correction for multiple comparisons (sleepy uncorrected p-value 0.022, corrected p-value 0.066; asleep uncorrected p-value 0.009, corrected p-value 0.052). At 9-months, 21 infants (46% of those contributing to the analysis of attentive state) participated in the first videos run, the first pictures run, and the second videos run (Figure 2.5F). Although the median interest decreased with each consecutive run (62.02%, 52.74% and 46.45% respectively), there were no significant differences in interest across run types. Indeed, there is some suggestion that the pictures might have been more engaging to 9-month-olds in general, as indicated by a substantial increase in the first quartile for the pictures as compared to either video run. Specifically, for 75% of runs, interest was greater than 37.67% for the first picture run, as compared to 16.96% for the first videos run or 13.12% for the second videos run. Similarly, the median fussing appeared to decrease slightly during the pictures task (6.46%, IQR = [0.0-24.5])

than the videos (11.97%, IQR = [2.73-32.06] for the first videos run). However, this difference was not found to be significant. Infants were sleepy for a significantly greater proportion of time in the second and third runs than in the first, suggesting that this may have been due to time in the scanner rather than in response to a particular stimulus.

2.5 Discussion

Through the FOUNDCOG project, we conducted a large longitudinal study using awake infant fMRI at two timepoints early and later in the first post-natal year, at 2 and 9 months of age, with the goal of characterising development of visual representations. In this work, we shared the factors that we considered when designing visual stimuli for this purpose, and metrics of subsequent data quality. Here, we reflect on these metrics in context of the broader literature and consider potential implications for the feasibility of adopting this technique.

The number of 2-month-olds who were willing to participate in an awake scan (n = 131; 97%), and the number of minutes acquired per participant (median 18.63 minutes, or 3.5 runs) were surprisingly high. The number of willing participants at 9 months (n = 65; 63%) and minutes acquired (median 13.02 minutes, or 2.5 runs) were also satisfactory. These acquisition rates exceeded our expectations based on the data available during the planning phase of the FOUNDCOG project. At the time, we anticipated that, across infancy, approximately 50% of experimental runs may be excluded due to motion, with a further 10–20% of timepoints excluded from analysis.

The comparative success of the younger cohort is consistent with the consensus in infant MRI more generally that 0–3-month-old infants are the easiest age group to scan. While this is typically attributed to the likelihood that these young infants will fall asleep (Hendrix and Thomason, 2022), we demonstrate that this intuition also extends to awake infants participating in a task-based scan. Although the field of awake infant fMRI remains in its infancy, a growing body of evidence supports the feasibility of this technique. A handful of studies have now been published, collectively scanning several hundred infants. Our findings are consistent with a recent analysis of data acquisition trends across many of these studies, which likewise found that younger infants tend to be more compliant with awake scanning than older infants, with participation decreasing linearly with age (Behm *et al.*, 2026).

When considering the acquisition rates reported here, it is important to note that, while the focus of this work is on the experimental paradigm, human factors are of the utmost importance in successful infant scanning. We share the measures taken to maximise infant and caregiver comfort in related work, detailing recruitment strategy, scan team composition, and hardware setup (King *et al.*, *in preparation*).

One of the prevailing challenges associated with infant fMRI is that even the most willing participants will inevitably move, as they cannot be asked to remain still. Several measures can be taken to minimise motion *a priori*; for example, infant comfort plays an important role (King *et al.*, *in preparation*). To minimise motion within each brain volume, we employed multiband acceleration, whereby multiple slices of the brain are acquired simultaneously, and each volume is therefore acquired more quickly. This resulted in a

short repetition time (TR) of 610 ms. Despite this clear advantage, multiband acquisitions also have disadvantages that have prevented their widespread adoption. These include poorer tissue contrast and distinct motion artefacts such as slice leakage, which refers to correlations between simultaneously acquired slices. However, these artefacts can be mitigated by retrospectively removing the affected timepoints (for example, by censoring them from the time series; Fitzgibbon *et al.*, 2020). Given the prevalence of motion in this cohort, we decided that the fast TR was worth this trade-off. Other awake infant scanning groups have typically utilised a single-band sequence, whereby volumes are acquired one slice at a time. Single-band sequences have longer TRs, ranging between 2–3 seconds in the studies published to date (Deen *et al.*, 2017; Ellis *et al.*, 2020). It is therefore necessary to be cognizant of TR when comparing motion distributions across studies and when choosing censoring thresholds, as TR sets the sampling frequency for framewise displacement (FD). We employed a threshold of 1.5 mm FD, which allowed us to retain a large percentage of our data (88% at 2 months and 92% at 9 months) and was relatively small compared to our voxel size (3 mm isotropic). Our FD threshold was more conservative than the 3 mm threshold employed by Ellis and colleagues, although their longer TR of 2 s should be noted (Ellis *et al.*, 2020). Both thresholds can be considered relatively liberal compared to that used by Deen and colleagues (0.5 mm, with a TR of 3 s; Deen *et al.*, 2017), reflecting a prioritisation of data quantity. Reassuringly, however, control analyses using more stringent motion thresholds have been found to produce similar patterns of results in analyses of the FOUNDCOG dataset to date (see Chapter 3 and O'Doherty *et al.*, 2026).

Motion correction methods for fMRI developed to date remain relatively limited. The majority of pipelines correct for rigid body motion (translation and rotation), using a registration method such as MCFLIRT, which outputs an estimate of FD (Jenkinson *et al.*, 2002). Many then use these FD estimates to implement some form of censoring or interpolation to remove the effects of motion corrupted volumes at a later stage of preprocessing (Power, Schlaggar and Petersen, 2015). These methods have been found to be sufficient for task-based MRI, with resultant improvements in data quality (Siegel *et al.*, 2014). While it is clear that infants move substantially more while awake than while asleep, the types of analyses conducted using functional activity acquired during sleep (resting-state functional activity) have been shown to be particularly susceptible to the effects of motion (Van Dijk, Sabuncu and Buckner, 2012; Power *et al.*, 2012; Power *et al.*, 2014). Thus, while motion is higher during awake task-based scanning, the effects are considered less extreme. However, motion during awake scans remains a key challenge, indeed task-based scanning has its own unique challenges such as the potential for stimulus-correlated motion (Hajnal *et al.*, 1994). At present, our toolkit for dealing with motion is still limited. Advances in motion correction techniques, such as prospective correction and deep learning-based methods enabled by advanced ground truth motion tracking, will therefore greatly enhance the viability of awake infant fMRI by increasing possible data yields.

The prevalence of motion in awake infant fMRI also influenced how regions of interest were defined. As awake infants move considerably more than sleeping infants or awake adults (Figure 2.3E), we opted not to acquire anatomical scans while infants were awake, as these longer acquisitions are particularly vulnerable to motion artefacts. Although a

high proportion of 2-month-olds (80%) and many 9-month-olds (53%) fell asleep during the session (Figure 2.3A), anatomical scans were unavailable for a number of participants who otherwise contributed high-quality functional data. To maximise the amount of usable functional data, regions of interest were therefore defined at the group level using adult atlas-based parcellations warped into infant space (Amunts et al., 2020b). This approach has clear limitations, as the resulting regions were not defined using infant functional data and may therefore only approximate the functional organisation of the developing ventral visual stream.

More broadly, progress is being made in functionally localising visual regions with awake infant fMRI. A modified retinotopy protocol has been successfully implemented in infants (Ellis et al., 2021b), enabling the delineation of subject-specific visual regions, and new approaches are being developed to localise regions on the basis of responses during movie watching (Ellis et al., 2025). Given the limited window of infant engagement, however, incorporating such paradigms into the awake scanning protocol would consume a substantial proportion of the total acquisition time. Thus, while adult template-based parcellation is imperfect, it currently offers a practical entry point for characterising the developing ventral visual stream. In particular, for group-level analyses, some consequences of imperfect regional definition may be mitigated by aggregating data across many participants and runs. Ongoing related work by the FOUNDCOG team is using the subset of infants from whom anatomical scans were acquired to further evaluate the validity of this approach.

At both ages, infants were found to attend to the display for the majority of scan time, looking away or closing their eyes for just a very short proportion of time. We also quantified whether the infant appeared to be interested in the stimulus. Interest was defined as a subset of the time when the infant was looking at the screen and was also scanning the stimulus with their eyes. Overall, 2-month-olds were more interested in the stimuli for a significantly greater proportion of the time than 9-month-olds (median proportion 75.27% and 48.52% respectively). The same pattern was observed in the subset of infants who tolerated awake scanning well and completed at least three runs. This indicates that although the level of interest expressed by the 9-month-olds was lower, it was sufficient and did not necessarily preclude continued participation. As previously mentioned, the videos were always presented before the pictures. Comparisons between the tasks must therefore be interpreted with caution as they will always be confounded by ordering effects. 2-month-olds appeared to be more interested in the videos than the pictures; this was in line with our expectation that the dynamic naturalistic stimuli would be more engaging. Interestingly, this was not found to be the case for the 9-month-olds, who expressed no preference for either stimulus set. In fact, there were some trends to suggest that, in general, the 9-month-olds may have found the pictures more interesting than the videos, although these trends did not reach significance. This is particularly intriguing as a non-trivial proportion of the pictures paradigm consists of a fixation cross, in contrast to the videos which consisted of a continuous stimulus or engaging attention getter.

When designing the pictures task, although we were hopeful that the looming would be helpful, we were sceptical that this task would sufficiently capture infant interest. The

success of this task was therefore very reassuring. The viability of presenting static pictures has since been validated in other work (Yates *et al.*, 2025). With hindsight, there are aspects of the implemented pictures paradigm that could be improved. One limitation of this task was that the range of the jittering of the fixation cross (0.5-1.5 s) was insufficient to enable a stimulus minus baseline comparison. This was because we were doubtful that infants would be content to wait through fixations of the length required for such a comparison. A longer fixation was therefore considered to be high risk, and we opted to prioritise maintaining infant engagement. Recent work has demonstrated that longer inter-trial intervals are indeed possible (Yates *et al.*, 2025). Yates and colleagues (2025) employed a jittered inter-trial interval of 2, 4 or 6 seconds. However, there were distinct differences in the stimulus presentation across each of these studies. Whereas our inter-trial interval comprised a static fixation, and the subsequent trial a looming object, Yates and colleagues (2025) utilised a dynamic (looming or receding) pattern during the intertrial interval, followed by a static stimulus image. Thus, it remains to be determined whether tolerance to longer inter-trial intervals would extend to a static fixation.

One of the goals of this work was to share our stimulus design process, not as we believe it to be optimal, but rather as it was found to be effective, and it serves to highlight the many considerations and trade-offs that one must make when designing a paradigm for awake infant fMRI. Accordingly, it might be of benefit for those planning similar studies. Our process can therefore be thought of as the route we chose to navigate these decisions, with the purpose of providing a sense of the landscape rather than dictating any particular path.

Having considered the general success of the acquisition, we now turn to reflect on the efficacy of our stimuli for achieving their specific goals, with reference to related work. The videos paradigm achieved a robust inter-subject correlation, with broad activation in both visual and frontal regions already observed by 2 months of age. As hoped, this rich data has enabled multiple analyses across diverse cognitive processes which are detailed in related publications (Santillo *et al.*, *submitted*).

The pictures, deemed *a priori* to be a high-risk stimulus choice based on the data available at the time, successfully achieved their ambitious goal of characterising rich group-level visual responses, as detailed in (O'Doherty *et al.*, 2026). Representational similarity analysis has only sparsely been applied to infant neuroimaging data (Emberson *et al.*, 2017b; Bayet *et al.*, 2020). In fMRI, it has been used to investigate broad category structure across high-level visual cortex, in response to four categories — faces, bodies, objects and scenes (Deen *et al.*, 2017; Kosakowski *et al.*, 2022). Similarly, using a larger dataset of 128 images, broad category structure in response to faces, bodies, toys, and houses has been investigated using EEG. Despite the relatively low-dimensional representational space (4x4), layer-wise comparison to DNNs revealed that the coarse representational structure across the visual cortex of infants was more like earlier layers of DNNs (Xie *et al.*, 2022). Thus, the strength of RSA to characterise high-dimensional representational spaces had not previously been utilised in infants. To our knowledge, the FOUNDCOG dataset therefore enabled the first demonstration of fine-grained object-level RSA in awake infant fMRI. The fine-grained characterisation of infant responses enabled the application of advanced computational models, specifically comparison with DNNs. We refer the interested reader to (O'Doherty *et al.*, 2026), where we report the findings from this

analysis, and to Chapter 3 where we compare to developmentally-inspired DNNs to further characterise the features contributing to such representations. Here we focus on the strengths and weaknesses of our dataset, with the goal of highlighting what it enabled us to achieve and what was or was not possible given our design.

Visual representations were calculated at the group level, taking the run as the unit of the representational similarity matrix (RSM). One challenge with infant data is that participants contribute a variable number of usable runs. To maximise the utility of each run, and to increase our degrees of freedom, we opted to compute across-subject RSMs using the run as the basic unit of analysis. Each RSM was calculated by comparing each subject–run pair with every other subject–run pair. This resulted in a very large number of RSMs, enabling robust group-level estimates with reliable bootstrapped confidence intervals. While this approach has clear advantages for noisy infant data, it does not fully exploit the strengths of RSA, which is typically applied at the individual level (Kriegeskorte, Mur and Bandettini, 2008). A variance partitioning analysis indicated that there was a substantial amount of noise at the individual level(O'Doherty *et al.*, 2026), highlighting a weakness of our data for questions that rely on individual differences.

This had implications for individual-level longitudinal comparisons between 2 and 9 months. For the subset of infants who contributed multiple usable picture runs at both ages (n = 36), we derived individual-level RSMs. When comparing RSMs across ages, infants' own RSMs at 2 and 9 months were not significantly more similar to each other than to the RSMs of other infants, see (O'Doherty *et al.*, 2026 for details). These analyses therefore highlight the limited ability of our dataset, in its current form, to support robust

inferences about individual differences in representational change over time. This limitation may also partly reflect the fact that regions of interest were defined using atlas-based parcellations in template space rather than subject-specific regions in native space, which is likely better suited to characterising average patterns of organisation than individual differences in representational geometry.

Given these weaknesses at the individual level, if the *a priori* goal were solely to characterise group-level representations, a multi-cohort cross-sectional design would be more economical. In such a case, our 2-month-old sample could be considered over-powered, although note that the noise ceiling does continue to increase modestly with increasing sample size (Figure 2.4A, C). For a longitudinal design, however, the current sample size is at the edge of viability, as the number of infants comparable at 2 and 9 months (for the pictures task) drops to 36 infants (Figure 2.4B, C). For the FOUNDCOG project, imaging took place in the context of a larger study which includes behavioural follow-up, thus the appropriate sample size for such comparisons remains to be determined (Linke *et al.*, 2018).

Our intent is not to discourage individual-level designs. Indeed, individual differences will likely be essential for unlocking the full potential of awake infant fMRI for both basic science and clinical goals (Nielsen, Graham and Sylvester, 2023; Olson *et al.*, 2025). Rather, we aim to highlight the necessity of recruiting large samples to obtain data from a sufficient number of infants, with a sufficient number of repetitions, and preparing for high rates of attrition for longitudinal analyses. Following which, as we have shown, the resulting data could still be noisy at the individual level. Opting for a design that also

addresses interesting questions at the group level would therefore be a strategic safeguard.

The tasks developed here, and the analyses they have enabled, constitute a significant methodological advance and serve to highlight the viability of awake infant fMRI at scale. Given the success of the pictures paradigm, an appealing future direction would be to scale to a larger set of objects, enabling even finer-grained characterisation of the VVS. A recent EEG study presented a series of 200 images (50 objects, each with 4 exemplars) to infants (n = 42; 2–12 months), leveraging innovations in temporally resolving stimulus responses recorded using EEG to present 5 images per second (Grootswagers *et al.*, 2025). MRI does not have the temporal resolution for such a rapid-fire presentation. However, given the level of infant interest in our paradigm, it seems viable that the number of conditions could be increased beyond what was implemented here. We also prioritised the acquisition of the videos relative to the pictures, such that the pictures comprised only 43% (1.5/3.5 runs) of awake scan time at 2 months and 40% (1/2.5 runs) at 9-months. Thus, there is scope to increase the number of stimuli. However, several uncertainties remain: for example, would the pictures task be as successful if it were always presented first, and would infants attend for the same amount of time in the absence of stimulus variability? Alternatively, one might prioritise a greater number of repetitions of a smaller set of stimuli to attempt to decrease noise at the individual level, although it is unclear how many repetitions of the same stimulus would be tolerated before infants lose attention. The scale of data collection that is currently viable in adults remains some way off for infants, however, the field is taking its first steps in that direction (Allen *et al.*, 2022).

Awake infant fMRI has unique potential to shed light on the principles of both the developing (Tansey *et al.*, 2024) and the adult brain (Ellis and Turk-Browne, 2018). Here we take initial steps towards bridging the gap between small-sample studies of awake, behaving infants and large-scale fMRI studies conducted with sleeping infants (Howell *et al.*, 2019; Eyre *et al.*, 2021; Nelson, Frankeberger and Chambers, 2024; Dean *et al.*, 2024). We believe that the data presented here, and by others (Behm *et al.*, 2026), provide compelling evidence that awake infant fMRI is viable at scale. However, as highlighted, the modality presents a unique convergence of challenges that require special consideration during experimental design. We hope that sharing our process for navigating these challenges, alongside our obtained success rates, and the accomplishments and nuances of the resulting data, will provide a useful and thought-provoking reference for researchers in the field considering adopting this challenging yet rewarding modality.

Chapter 3

The Development of Tuning to Coarse and Fine Visual Features in Infant Visual Cortex

Objects can be discriminated using visual features across a range of spatial scales, from global shape to fine detail. In adults, the ventral visual stream represents object features across these scales. How such tuning develops in infancy remains poorly understood: are infant visual representations dominated by coarse structure, as might be expected given the limited acuity of infants' early visual experience, or do they already include fine detail? We used a NeuroAI approach combining awake infant fMRI with deep neural networks (DNNs) as developmentally inspired computational models to characterise developing object representations. Awake fMRI responses were acquired from 2-month-old (n = 130) and 9-month-old (n = 65) infants, and adults (n = 17). DNNs were trained on object images with graded amounts of blur to systematically vary sensitivity to coarser versus finer object features. Comparing model and brain responses revealed that tuning across spatial scales in V1 was already comparable to adults by 2 months, whereas downstream visual regions showed protracted development and remained immature at 9 months. Across development, feature complexity increased along the visual hierarchy, with downstream regions best captured by later DNN layers. As young as 2 months, V1 responses were better explained by category-trained DNNs than by control networks that were untrained or trained on synthetic retinal waves simulating in utero activity prior to eye opening. Together, these findings suggest that infant visual cortex has access to rich object-specific features across a range of spatial scales by 2 months and place new empirical constraints on theories of how coarse and fine object representations develop in the human visual system.

3.1 Publication details

This chapter is in preparation for publication with contributions from the following co-authors.

3.1.1 Authors and affiliations

Áine T. Dineen[1,2], Clíona O'Doherty[3], Anna Truzzi[4], Anna Kravchenko[1,2], Alex R. Wade[5], Graham King[12], Enna-Louise D'Arcy[1,2], Angela T. Byrne[6,7], Ailbhe Tarrant[8,9], Adrienne Foran[10,11], Eleanor J. Molloy[12-15], Rhodri Cusack[1,2].

1. *Trinity College Institute of Neuroscience, Trinity College Dublin, Dublin, Ireland.*
2. *School of Psychology, Trinity College Dublin, Dublin, Ireland.*
3. *Department of Psychology, Stanford University, CA, United States of America.*
4. *School of Psychology, Queen's University Belfast, Northern Ireland, United Kingdom.*
5. *Department of Psychology, University of York, England, United Kingdom.*
6. *Radiology, The Coombe Hospital, Cork Street, Dublin 8, Ireland.*
7. *Radiology, Children's Health Ireland, Dublin, Ireland.*
8. *Radiology, The Rotunda Hospital, Parnell Square, Dublin 1, Ireland.*
9. *Radiology, Children's Health Ireland at Temple Street, Dublin 1, Ireland.*
10. *Neonatology, The Rotunda Hospital, Parnell Square, Dublin 1, Ireland.*
11. *Children's Health Ireland, Dublin, Ireland.*
12. *Discipline of Paediatrics, Trinity College Dublin, the University of Dublin, Dublin, Ireland.*
13. *Trinity Translational Medicine Institute, St James Hospital & Trinity Research in Childhood Centre, Dublin, Ireland.*
14. *Neonatology & Neurodisability, Children's Health Ireland, Dublin, Ireland.*
15. *Paediatrics, Coombe Women's and Infant's University Hospital, Dublin, Ireland.*

3.1.2 Author contributions

Conceptualisation: **Á.T.D.**, C.O'D., A.Tr., A.K., R.C.

Methodology: **Á.T.D.**, C.O'D., A.Tr., A.K., R.C.

Recruitment and resources: **Á.T.D.**, C.O'D., A.Tr., A.K., G.K., E.L.D., A.Ta., A.T.B., A.F., E.J.M., R.C.

Investigation: **Á.T.D.**, C.O'D., R.C.

Ethics and project administration: **Á.T.D.**, C.O'D., A.Tr., G.K., A.Ta., A.T.B., A.F., E.J.M., R.C.

Data curation, preprocessing, and software: **Á.T.D.**, C.O'D., A.Tr., G.K., R.C.

Formal analysis and validation: **Á.T.D.**, R.C.

Writing – original draft: **Á.T.D.**, R.C.

Writing – review & editing: **Á.T.D.**, C.O'D., A.Tr., A.R.W., R.C.

Visualisation: **Á.T.D.**, R.C.

Supervision: E.J.M., R.C.

Funding acquisition: R.C.

3.2 Introduction

When an infant sees an object, what features does their visual cortex encode? In the first months after birth, the eyes mature rapidly, such that the ability to see fine details improves (Banks and Bennett, 1988; Hendrickson and Boothe, 1976) and sensitivity to contrast increases (Braddick and Atkinson, 2011). These perceptual abilities, known as visual acuity and contrast sensitivity, respectively, can be estimated by presenting infants with simple patterns, such as flickering, striped gratings, and recording the resulting modulation of electrical potentials from the scalp. These visually evoked potentials reflect activity in the primary visual cortex (V1; Norcia and Tyler, 1985; Norcia, Tyler and Hamer, 1990), and indicate that, even early in infancy, basic, low-level visual features are being processed by the cortex.

Objects, however, are far more complex than the simple stimuli used in these paradigms (Kayser, Körding and König, 2004). They are composed of features across a range of spatial scales, from fine details to global shape. When the brain processes an object, beyond detecting isolated low-level features, it must also process information across these spatial scales. A cat, for example, has fine-grained features such as fur texture, but also coarser features with a canonical shape, such as pointy ears. In adults, objects are processed by a series of cortical regions collectively known as the ventral visual stream (VVS). The VVS represents visual features across a range of complexities: earlier regions tend to process simple perceptual features, intermediate regions process increasingly complex patterns such as contours and textures, and higher regions process more

abstract, semantic information such as object category (Mishkin, Ungerleider and Macko, 1983; Goodale and Milner, 1992; Downing *et al.*, 2006).

While it is intuitive to consider familiar object features such as a cat's fur or pointy ears, neuronal tuning along the VVS is often more complex and abstract than such tangible descriptors (DiCarlo and Cox, 2007). Deep neural network models (DNNs) have proven highly effective at characterising visual responses along the VVS (Kriegeskorte, 2015), and demonstrate particular value for characterising the types of non-intuitive, intermediate-level complexity features encoded by the region (Long, Yu and Konkle, 2018; Jagadeesh and Gardner, 2022; Ayzenberg and Behrmann, 2022; Henderson, Tarr and Wehbe, 2023). Trained on object classification tasks, DNNs learn features that are informative for object recognition (Yamins and DiCarlo, 2016). Like the regions of the VVS, DNNs comprise a series of hierarchical layers that encode increasingly complex features (Güçlü and van Gerven, 2015; Cichy *et al.*, 2016). In adults, the representations learned by these models align well with object representations along the VVS (Schrimpf *et al.*, 2018), and are widely regarded as the current best models of its processing (Storrs *et al.*, 2021), albeit with some limitations (Bowers *et al.*, 2022). Comparing the features learned by DNNs to those encoded by the brain – an approach often referred to as NeuroAI – has driven substantial progress in modelling adult visual cortex, generating novel insights and hypotheses about the organisation of the VVS, particularly in higher areas (Kriegeskorte, Mur and Bandettini, 2008; Doerig *et al.*, 2023; Bao *et al.*, 2020). Although the features extracted by these models can be difficult to interpret directly, this can also be considered a strength, as they capture statistical structure beyond our current conceptual vocabulary. Further, innovative visualization methods exist and can

be used to reveal aspects of their tuning, providing tangible insights (e.g. Görgün, Schiele and Fischer, 2025). An additional advantage is that their training can be systematically manipulated, for example by changing the statistics of the input data. This creates different models that can be compared to the brain to determine which best aligns with neural responses, providing insight into the features that the brain encodes (Kietzmann, McClure and Kriegeskorte, 2019). Given the success of these models at capturing responses along the adult VVS, they are excellent candidates for probing the features encoded by the developing visual system. We have recently validated that such models align with infant visual responses (O'Doherty *et al.*, 2026), motivating the use of advanced computational models for a more thorough investigation of the features driving developing object responses.

How the visual cortex develops the ability to process the features that comprise objects remains poorly understood (Ayzenberg and Behrmann, 2024). One possibility is that object tuning in V1 matures early, alongside many other key functions of this region (Kiorpes, 2016). Even before birth, spontaneous patterns of neuronal activity from the retina, termed retinal waves, shape the developing visual system (Huberman, Feller and Chapman, 2008). Retinal waves help instantiate tuning to low-level visual features in V1, by establishing basic receptive field properties such as retinotopy and orientation selectivity (Chapman, Gödecke and Bonhoeffer, 1999; Huberman, Feller and Chapman, 2008). In contrast, subsequent regions in the VVS lag behind V1, both functionally (Kiorpes, 2016) and anatomically, raising the possibility that tuning to object features might emerge first in V1 and only later in downstream extrastriate regions (V2–V4). Consistent with this idea, computational modelling has suggested that activation

patterns in V1 could help drive the subsequent development of tuning in regions such as V2 (Imam and L Finlay, 2020).

An alternative hypothesis is that tuning to object features develops in parallel across the processing hierarchy, despite differences in maturation of anatomy and single cell physiology across regions. When DNNs are trained to recognise objects, representations at multiple layers are jointly optimised, such that early, intermediate and late stages all become tuned to task-relevant features at approximately the same time (Vogelsang *et al.*, 2018; Lu *et al.*, 2025a). By analogy, the biological visual system might also learn about object features via experience-dependent optimisation, simultaneously fine-tuning V1 and subsequent regions. Early visual experience plays an important role in driving the development of typical responses in V1, although it is unclear whether it fine-tunes receptive fields in the region or simply maintains tuning established by innate mechanisms (Hirsch and Spinelli, 1970; Heitmann *et al.*, 2023; Kiorpes, 2015).

A third hypothesis is that the development of multiscale object feature tuning follows a non-hierarchical trajectory relative to V1. This has recently been shown for the dorsal stream with intermediate-level region MT developing prior to upstream regions (Mundinano, Kwan and Bourne, 2015).

Even in adults, our understanding of how V1 represents complex stimuli such as objects remains incomplete. In infancy, while single-cell response properties in V1 mature relatively early in non-human primates, population-level responses and tuning to complex, naturalistic stimuli have been sparsely studied (Kiorpes, 2015). We have

recently shown that, relative to subsequent regions, category-selective responses in V1 are comparatively mature by 2 months of age, although they are not yet fully adult-like (O'Doherty *et al.*, 2026). Indeed, infant V1 is known to remain immature in several respects: its microanatomy is not yet fully developed (Siu and Murphy, 2018), and its recurrent circuitry – horizontal and feedback connections – continues to undergo protracted maturation. Horizontal connections within V1 emerge prenatally and support rudimentary contour integration by around 3 months (Taylor *et al.*, 2014), but do not reach adult-like organization until around 2 years of age, nor do feedback projections from V2 (Burkhalter, 1993). Such recurrent connections are increasingly recognised as important contributors to processing naturalistic objects (Uyar *et al.*, 2016; Lamme, 2014; Petro, Vizioli and Muckli, 2014). These anatomical and physiological immaturities therefore suggest that multiscale object feature tuning in V1 may continue to develop across infancy.

This view aligns with a broader reconceptualization of V1 as a computationally rich region, rather than purely a simple edge detector. Classical models characterised V1 as a bank of Gabor filters tuned to orientation and spatial frequency (Hubel and Wiesel, 1962). While this remains a predominant and important function of the region, it is no longer considered its exclusive role (Rust and Movshon, 2005), recurrent and feedback connections within and to V1 are now understood to contribute to its responses to complex stimuli such as natural objects (Lee *et al.*, 1997; Petro, Vizioli and Muckli, 2014; Kietzmann *et al.*, 2019). In particular, at the timescale of fMRI, responses in V1 will contain contributions from such recurrent connections, which can be modelled either with recurrent deep neural networks or with deeper layers of standard feedforward

networks that serve as static approximations to recurrent processing (van Bergen and Kriegeskorte, 2020). There is also evidence that even the initial feedforward sweep in V1 may be more complex than previously appreciated. Single-cell recordings from mature macaques show that, in response to naturalistic stimuli such as objects, features from intermediate layers of deep convolutional networks provide superior models of V1 responses compared to traditional filter-bank models (Cadena *et al.*, 2019). Moreover, even when controlling for model complexity and receptive field scale across architectures, V1 responses are still best captured by early-to-intermediate convolutional layers, rather than the very first convolutional layer (Miao and Tong, 2024), suggesting an important contribution of multiple nonlinear processing stages to modelling its responses. Such intermediate layers tend to respond to a mixture of low and intermediate complexity features, from colour to object parts (Bau *et al.*, 2020). Given this potential complexity, and the protracted maturation of recurrent circuitry, it is plausible that multiscale object feature tuning in V1 continues to develop across infancy.

To characterise the emergence of object feature tuning across spatial scales, and subsequently distinguish these hypotheses, we took a NeuroAI approach, combining awake infant fMRI with computational modelling using DNNs. Awake infant fMRI (Deen *et al.*, 2017; Ellis *et al.*, 2020; Yates *et al.*, 2025) is yielding new insights into the development of the VVS (Deen *et al.*, 2017; Ellis *et al.*, 2021b; Kosakowski *et al.*, 2022; Ellis *et al.*, 2025; O'Doherty *et al.*, 2026), for example, revealing that retinotopic organization is present as early as 5 months in humans (Ellis *et al.*, 2021b; Ellis *et al.*, 2025), similar to infant macaques (Arcaro and Livingstone, 2017). Through the

Foundations of Cognition (FOUNDCOG) project, we measured developing visual responses over the first year by conducting a large longitudinal study using awake fMRI. Functional brain activity was acquired from infants while viewing visual stimuli at 2 months (n = 130), when acuity is still quite poor (Norcia and Tyler, 1985; Banks and Bennett, 1988) and the cumulative amount and variety of visual experience has been limited, and at 9 months (n = 65), when acuity, relative to our stimulus display, is near adult levels, and infants have amassed substantially more cumulative and varied visual experience (Jayaraman, Fausey and Smith, 2015; Smith *et al.*, 2018; Long *et al.*, 2024). Many of these infants viewed a series of objects, with 101 2-month-olds and 44 9-month-olds providing high-quality, usable data, allowing group-level visual responses to be characterised. Using these data, we have recently shown that infants already have rich category representations along the VVS at 2 months that strengthen with age (O'Doherty *et al.*, 2026). Here, we leverage this dataset to probe the features driving responses to objects throughout development by modifying the visual experience of DNNs such that they learn to extract different stimulus features.

As described above, object recognition involves processing features across a range of spatial scales. While DNNs can, in principle, learn to use features at any scale, standard training regimes for object classification bias models towards the use of local image features (Geirhos *et al.*, 2018a; Jang and Tong, 2021). Humans, in contrast, are biased towards object classification based on shape from an early age (Petras et al., 2019; Landau, Smith and Jones, 1988). Blurring an object emphasises global shape while suppressing fine detail. Training DNNs on blurry images encourages the development of more shape-based representations (Vogelsang *et al.*, 2018; Avberšek, Zeman and Op de

Beeck, 2021), changing the object features encoded by DNNs. For example, blurry training increases receptive field sizes in early network layers (Vogelsang *et al.*, 2018). Blurry-trained models have been found to show superior alignment with visual responses recorded from mouse V1 and the VVS in monkeys and human adults (Li et al., 2023; Jang and Tong, 2024). This pattern has been replicated across a diverse range of DNN architectures, from convolutional neural networks (CNNs) to vision transformers (Jang and Tong, 2024; Lu et al., 2025b).

To characterise visual object representations across a range of spatial scales, we trained DNNs on a large and diverse set of objects (Deng et al., 2009) to which varying amounts of blur were applied. We used a ResNet-50 (He et al., 2015), a widely used convolutional neural network architecture that has been extensively employed in studies of visual object recognition and model–brain correspondence in the VVS. These blurry-trained models were then presented with the same set of objects as the infants, also blurred by varying amounts. This allowed representational similarity analysis (RSA; Kriegeskorte, Mur and Bandettini, 2008) to be used to characterise and compare model and brain responses (Figure 3.1B), enabling us to infer the features that contributed to object processing in the developing visual cortex across spatial scales. By systematically varying the amount of blur applied both to the images used to train the DNNs and to the test stimuli used for RSA, we could separate two factors: training blur, which modified the spatial scale of object features to which the DNN was most sensitive, and test blur, which emulated the resolution of the input received by a given region at a given test age (2 or 9 months). For each combination of train and test blur, the network layer that maximised correspondence with a given region for a given age was selected to capture the most

appropriate level of feature complexity (Figure 3.1C). This gave a single value for each combination of train and test blur, creating a two-dimensional tuning map indicating tuning to spatial scale for each age and region (Figure 3.1D). This approach enabled us to test whether the visual cortex is tuned to visual object features across spatial scales by 2 months, how such tuning develops over the first year and into adulthood, and whether its developmental trajectory follows a V1-first, parallel, or a non-hierarchical, region-specific pattern. Developing visual responses were also compared to a model trained on simulated retinal waves to test whether tuning in visual cortex was object specific or could instead be explained by simpler visual statistics.

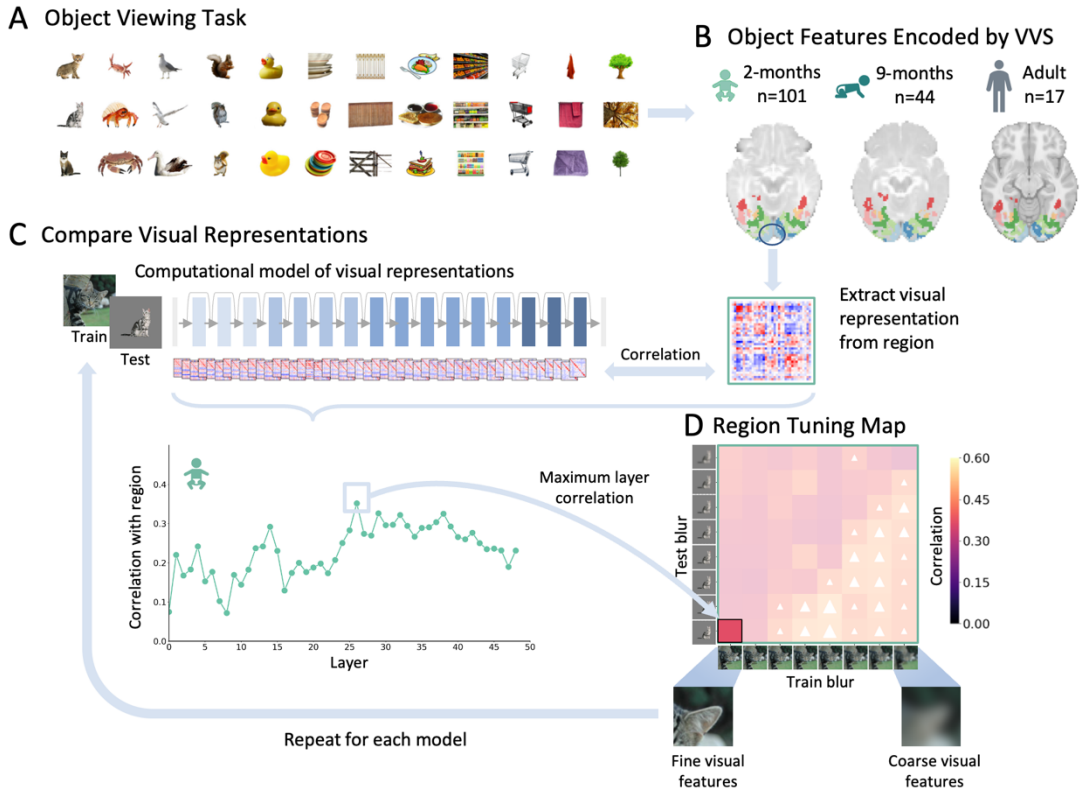

Figure 3.1 Object feature tuning was characterised along the developing ventral visual stream using awake infant fMRI and deep neural network modelling.

Participants viewed 36 object images (3 exemplars from 12 categories spanning animate, inanimate large, and inanimate small objects) (A). Functional responses to these test stimuli were recorded using fMRI. For each region along the ventral visual stream, group-level visual representations were derived for 2-month-olds (n = 101), 9-month-olds (n = 44), and adults (n = 17) by computing pairwise response similarities across stimuli and averaging across participants (B). The same images were presented to a trained deep neural network (DNN) and layer-wise activation patterns were extracted to generate corresponding model representations. Model–brain correspondence was quantified by correlating the similarity structure of each model layer with that of each brain region and age group. The layer yielding the highest correspondence was taken as the visual representation that best captured the features encoded by a given region, allowing the appropriate level of complexity to be captured (C). To assess tuning to visual features across spatial scales, DNNs were trained and tested on images with varying levels of blur (train blur, x-axis; test blur, y-axis). Systematically varying blur enabled estimation of the spatial scales to which each region was tuned at each developmental stage (D).

3.3 Results

3.3.1 Object responses across adult early visual cortex are best modelled by deep neural networks trained on blurry images

Deep neural networks (DNNs) were trained on images of objects with graded levels of blur. As hypothesised, blurrier trained models exhibited a preference for object features with progressively lower spatial frequencies (see Methods section 3.5.4.B; Appendix 3, Figure A3.3), allowing them to be used to systematically probe the spatial scale of developing object responses (Figure 3.1D).

We first validated our approach by characterising multiscale object feature tuning along the adult VVS. A multiscale tuning map was calculated (as described in Figure 3.1) for the following cortical regions (defined using Jülich atlas parcellations, see Methods section 3.5.5.A): the early visual cortex (EVC; V1–V3), intermediate regions (V4, VO1) and a higher category region (FG4).

In adults, regions comprising the EVC (V1–V3) and intermediate region V4, were best modelled by training on intermediate to high levels of blur and testing on sharper images (Figure 3.2A) and were typically best fit by intermediate model layers (Figure 3.2F). In contrast, the sharpest trained models, those that are typically used for model–brain comparisons, were consistently among the poorest models of these regions. The striking asymmetry of the tuning patterns demonstrates that it is specifically the presence of fine details during model training that decreases representational correspondence, while at test time, the presence of such details is complementary to representations learned using coarser features, increasing representational correspondence. This expands previous findings that training on blurry images can increase model alignment with mature brain representations (Jang and Tong, 2024; Li *et al.*, 2023) by highlighting the complementary influence of fine details at test time to better characterise responses in early regions. Thus, in the EVC and V4, both coarse and fine features make an important contribution to visual responses. These findings indicate that DNNs must learn sensitivity to coarser features through blurry experience during training, while, even without experience, finer features contribute to visual representations in a manner that increases alignment with the EVC.

Tuning patterns from later VVS regions – an intermediate region (VO1) and a higher-level region in the fusiform gyrus which overlaps with category selective cortex (FG4; Weiner *et al.*, 2017) – were distinct from those of early visual regions (Appendix 3, Figure A3.6 and A3.7). Consistent with the perceptual to semantic transition along the VVS, responses in these later VVS regions were best captured by later model layers (Figure 3.2*F*). Further, the visual representations in these regions were most similar to train-test combinations that resulted in the highest object classification accuracy (Appendix 3, Figure A3.8), rather than those that were trained and tested on features of particular spatial scales, suggesting that responses in these regions were driven by the category of the objects rather than the spatial scale of their constituent features.

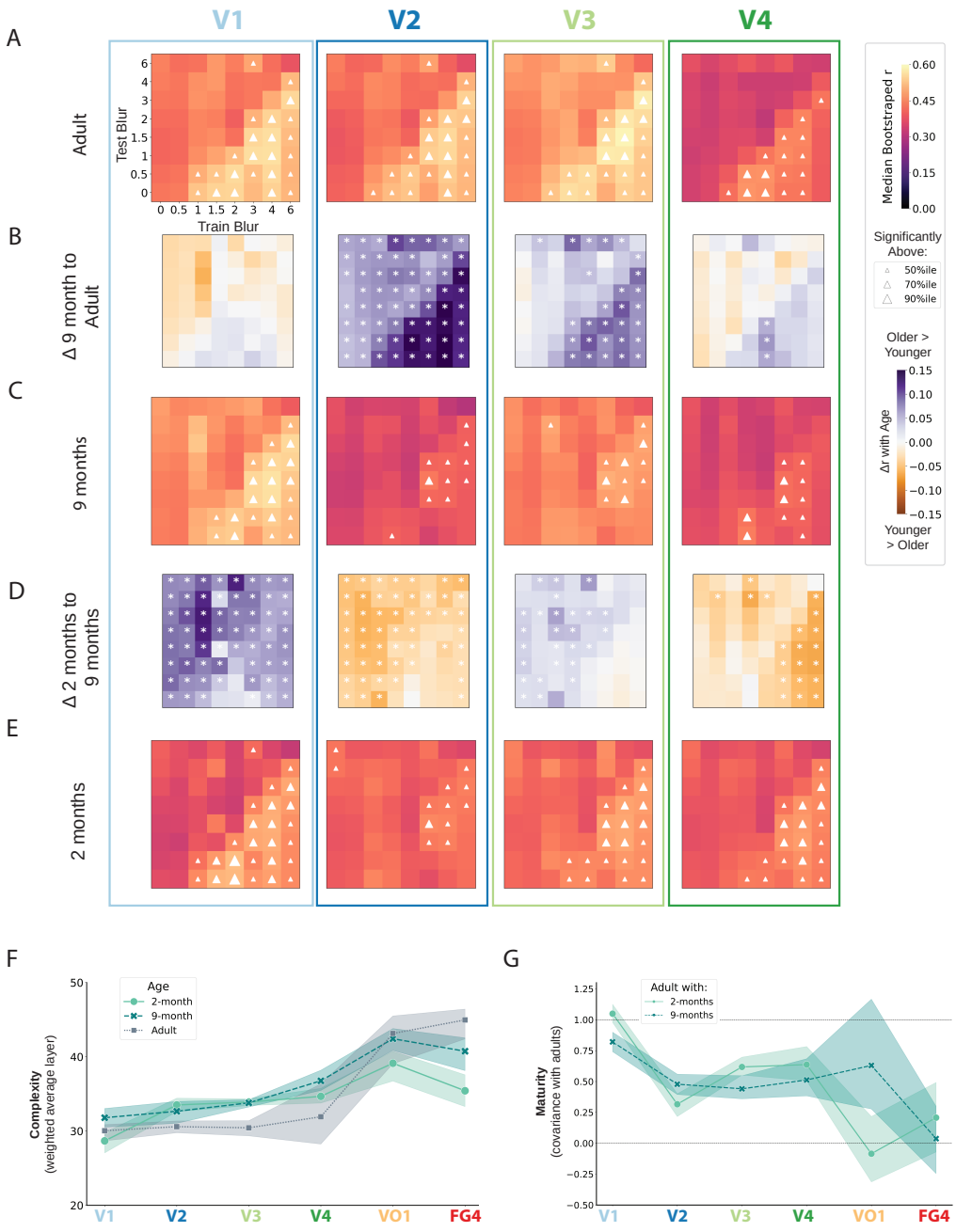

Figure 3.2 Development of multiscale object feature tuning along the ventral visual stream.

The pattern of tuning in V1 was already mature by 2 months, whereas tuning in later regions showed protracted development between 9 months and adulthood. Tuning to visual object features across spatial scales was assessed along the developing ventral visual stream as described in Figure 3.1 by training deep neural networks (DNNs) on objects with varying levels of Gaussian blur (train blur, x-axis; σ = 0-6) and testing across the same range (test blur, y-axis) for adults (A), 9-month-

olds (C), and 2-month-olds (E). For each cell in the tuning maps, model–brain correlations were computed for 1000 bootstrap resamples across subject-run pairs; each cell shows the median of the corresponding bootstrapped distribution. For a given tuning map, the best-performing models across all train-test blur combinations are indicated by triangles. A triangle marks a train–test blur combination whose model–brain correlation is significantly above a given bootstrapped percentile (50th, 70th, 90th) when compared to the full set of train-test combinations; larger triangles indicate significance at higher percentiles. Equivalent tuning maps, scaled by the corresponding noise ceiling for each age and region, are shown in Appendix 3, Figure A3.9. An alternative line-plot visualisation of tuning, showing the associated 95% confidence intervals for each train–test combination, is provided in Appendix 3, Figure A3.10. Difference maps were obtained by subtracting pairs of tuning maps with equivalent train (x-axis) and test (y-axis) blur values. The younger age was subtracted from the older (purple indicates higher correlations at the older age; orange indicates higher correlations at the younger age), with hue intensity indicating effect magnitude. Asterisks indicate significant age differences in the magnitude of model–brain correlations for a given train–test combination (B, D). The complexity of the features encoded by each region was estimated by calculating the weighted average layer for each bootstrap sample of the visual representations. The complexity is summarised as the median bootstrapped weighted average layer. Error bands denote 95% confidence intervals (CIs) derived from bootstrap resampling across subject-run pairs (F). Maturity of tuning was quantified by calculating, for each bootstrap, the covariance between infant and adult tuning maps, scaled by the variance of the corresponding adult region. The median bootstrapped value is plotted, and error bands denote the 95% CI (G).

3.3.2 Infant V1 is tuned to multiscale object features by 2 months

We first validated that the ResNet-50 architecture showed correspondence with the developing VVS, consistent with our previous findings using AlexNet (O'Doherty et al., 2026). To do so, we compared the layer-wise representations of a standard sharp-trained ResNet-50 with responses along the developing VVS (Appendix 3, Figure A3.4). The resulting pattern of correspondence was consistent with our previous findings, with trained models capturing significant components of visual responses across the VVS throughout development. Having validated the model architecture, we then applied our multiscale tuning approach to characterise the development of object feature tuning across spatial scales in infancy. Multiscale tuning maps were calculated for infants using the same procedure as for adults (Figure 3.2C,E). To quantify maturity, we calculated a scaled covariance measure (β, as defined in Methods section 3.5.6.E). Covariance was chosen because it is sensitive both to the qualitative pattern across the tuning map and to the magnitude of model–brain correspondence. To express maturity relative to adults, the infant covariance was divided by the covariance of the corresponding adult region, yielding a value that reflected the proportion of adult-like tuning. Values closer to 1 indicate more adult-like tuning patterns, whereas lower values indicate less mature tuning (Figure 3.2G).

Tuning patterns in V1 were found to be remarkably mature by 2 months, already exhibiting the same patterns as adults (Figure 3.2E). In V1, scaled covariance between 2-month-olds and adults was close to its highest expected value of 1 (median b across bootstraps = 1.049, CI = [0.978,1.125]). Between 2 and 9 months there was a general

increase in the magnitude of the correspondence between the models and V1 across the tuning map, that was not specific to a particular subset of train-test combinations. The pattern was otherwise consistent (Figure 3.2C-D). This lack of specificity suggests that the strengthening is not specific to a given spatial scale but rather may reflect a general strengthening of the neural response. Interestingly, this is consistent with the timeline of synaptogenesis in humans, with synaptic density in V1 rapidly increasing from around 2 months and peaking at around 8 months of age (Huttenlocher *et al.*, 1982). Importantly, the maturity of object responses in V1, including for sharp test images, suggests that neither visual acuity at the level of the eye, nor at the level of V1, is limiting the features to which the developing visual system has access. This partially reflects the choice of stimuli in our experiment, which were designed to be large and visible by 2-month-olds to minimise the influence of limitations in test-time acuity on the recorded object responses. This suggests that V1 was processing the full range of features present in the stimuli by 2 months of age, making full use of the available information across spatial scales. Like adults, infant responses were best modelled by intermediate model layers.

To investigate whether the cumulative amount of early visual experience affected the development of object feature tuning in V1, we leveraged data from a special subgroup of infants who participated in the FOUNDCOG study who were born preterm and had spent time in the neonatal intensive care unit shortly after birth (NICU group). Tuning maps were calculated for this group (n = 17) and compared to bootstrapped samples of controls who had participated in the same number of runs (run-matched controls; see Methods section 3.5.7). At the time of the scan, both groups were approximately 2

months corrected gestational age (CGA). However, because the NICU group were born prematurely, they had, on average, an additional 8 weeks of postnatal visual experience (median 7.71 weeks, IQR = [7.29-9.71]). Tuning in V1 was found to be consistent across both groups with no qualitative differences between the tuning maps (Appendix 3, Figure A3.11). This indicated that neither cumulative visual experience nor the atypical visual experience of the NICU (Marchal *et al.*, 2021) had an effect on the pattern of tuning in V1 at 2 months CGA.

3.3.3 Protracted development of object feature tuning in extrastriate visual cortex

In adults, tuning in extrastriate visual areas (V2, V3 and V4) closely resembled that of V1; however, this was not the case for infants (Figure 3.2A–E,G). In contrast to V1, which exhibited a mature (adult-like) tuning pattern by 2 months, object tuning in later VVS regions remained immature through 9 months of age, as indicated by scaled covariance with adults. This supports a hierarchical pattern of development of object feature tuning relative to V1 (Figure 3.2C,E). A strengthening of tuning in the train-test combinations of blur that best modelled adult visual representations did not emerge until after 9 months of age in V2 and V3, constituting a substantial lag from that of V1 (Figure 3.2D). Within each of these regions, models trained on high levels of blur and tested on intermediate levels of blur tended to be the best fitting models at 2 and 9 months of age, while high-resolution testing did not appear to be as good a fit for infants as it was for adults. The pattern of development observed in V4 was somewhat different. At 2 months, models trained on the highest amount of blur were among the best models of the region. This was not the case for older ages. Correspondence with these models was found to be

attenuated by 9 months, suggesting that V4 might be encoding coarser features in the youngest infants. Between 9 months and adulthood, a build-up of adult-like tuning was observed, as in V2 and V3. These trends hint at a possible tendency for extrastriate visual regions to be biased towards coarser object features throughout infancy, however, they must be interpreted with caution as evidence for this trend was not necessarily recapitulated in the difference maps between ages (Figure 3.2B,D). The maturity metric (scaled covariance; Figure 3.2G) reinforced that the overall tuning patterns of extrastriate visual areas were comparatively immature relative to V1, and that adult-like tuning primarily developed after 9 months of age. As with V1, intermediate layers tended to be the best models of extrastriate visual areas across age groups (Figure 3.2F).

To test whether differential patterns of motion might be contributing to the differences observed between infants and adults, the analysis was replicated using a low-motion subset of infant runs (see Methods section 3.5.6.F). A similar pattern of results was observed, suggesting that residual motion is not driving the observed effects (Appendix 3, Figure A3.12 and A3.13).

As in V1, no qualitative differences were observed between the NICU and control groups in extrastriate regions, suggesting that an additional 2 months of postnatal visual experience did not affect tuning patterns in these regions. In V3 and V4, however, the split-half noise ceiling was higher in the NICU group than in the bootstrapped controls, suggesting that response patterns were more similar across individuals in the NICU group than in the control group in these regions (Appendix 3, Figure A3.11).

Later visual regions, VO1 and FG4, showed progressive refinement of multiscale tuning across development. At 2 months, although visual representations in VO1 corresponded with model representations (Appendix 3, Figures A3.4 and A3.6), tuning remained immature. This immaturity was characterised by a lack of specificity, as the tuning pattern did not preferentially correspond to any particular train–test blur combinations. In addition, VO1 was best captured by earlier model layers than at older ages (Figure 3.2F). By 9 months, tuning in VO1 remained immature but was more adult-like: different train–test blur combinations were differentially effective at capturing the pattern of response, the region was better captured by later model layers, and correspondence was stronger for train–test blur combinations associated with higher object-classification accuracy (Figures A3.8, A3.6, and 3.2G). This developmental pattern is consistent with an account in which neural tuning becomes progressively more refined over development.

A similar pattern was observed in FG4, although 9-month-olds appeared less mature in this region than in VO1, consistent with hierarchical refinement of tuning along the VVS. These findings are also consistent with our previous work showing that category structure is already present in ventral temporal cortex at 2 months and strengthens progressively from 9 months to adulthood (O'Doherty et al., 2026).

3.3.4 What are the possible origins of object tuning in V1?

What could be the developmental origin of the remarkably mature object tuning observed in V1 by 2 months of age? Is experience with the visual world required or are

intrinsic properties of the infant brain and body, such as retinal waves, sufficient? As detailed, retinal waves play a role in setting up V1 tuning prior to the onset of visual experience, with basic retinotopy and orientation selectivity in place prior to eye opening (Chapman, Gödecke and Bonhoeffer, 1999; Huberman, Feller and Chapman, 2008). DNNs trained on classes of retinal waves have been shown to exhibit V1-like receptive field morphology (Cappell *et al.*, 2024; May, Dauphin and Gjorgjieva, 2025). Could the receptive fields learned through experience with such patterns of spontaneous activity be sufficient to account for the tuning that we see in V1? If V1 has developed basic receptive field properties by 2 months but does not exhibit specific sensitivity to object features, the representational geometry of a DNN trained on retinal waves in response to our test stimuli could be sufficient to account for the responses that we observe in infant visual cortex. To investigate whether this is the case, infant visual responses were compared to those of a DNN trained on classes of retinal waves (Figure 3.3A; Cappell *et al.*, 2024; Cappell and Stoll, 2023), and to an equivalent network trained on objects. If retinal waves are sufficient, both conditions would be expected to have a similar magnitude of model–brain correspondence. Alternatively, experience with retinal waves may be beneficial but experience with objects may also be required to capture responses in early visual cortex. In this case, increased model–brain correspondence would be expected for models pretrained on retinal waves and subsequently trained on objects. To test this, responses were also compared to a model that was pretrained on retinal waves and subsequently fine-tuned on objects. To control for effects of architecture, responses were also compared to an untrained network.

The retinal wave trained model was found to be a significantly inferior model of V1 than the object trained model across all ages when testing on sharp or blurry stimuli (Figure 3.3B-C), indicating that V1 has already developed specific sensitivity to object features by 2 months. For standard test stimuli, pretraining on retinal waves was found not to confer any benefit over object training alone (Figure 3.3B). Interestingly, when presented with blurry images, however, the model pretrained on retinal waves and fine-tuned on sharp objects did confer enhanced model–brain correspondence for both infants and adults relative to training on objects alone (Figure 3.3C). This suggests that, perhaps somewhat analogous to blurry training, experience with retinal waves modified the tuning of the receptive fields of the DNNs in a human-like manner.

As the retinal wave stimuli are greyscale, one possibility is that the ability to exploit colour could be driving the higher model–brain correspondence for the object trained model compared to the retinal-wave-trained model. To test this, RSA was replicated using greyscale versions of the FOUNDCOG test stimuli. The greyscale images yielded a similar pattern of results (Appendix 3, Figure A3.14), verifying that the difference between the retinal wave and object trained models reflects learned object-specific features that were not simply driven by image colour.

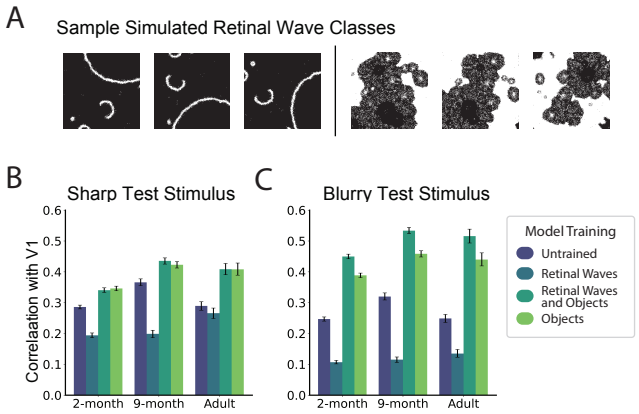

Figure 3.3 Features captured by infant V1 reflect object-specific tuning.

To test whether early V1 representations could be explained by features learned from retinal waves, patterns of spontaneous activity emanating from the retina that propagate through the visual system before the onset of sight, object responses were compared to deep neural networks that were: trained on simulated retinal waves (blue), trained on objects (light green), pretrained on retinal waves and fine-tuned on objects (dark green) and an untrained control (purple). The models were trained and publicly released by Cappell and colleagues (Cappell and Stoll, 2023). Sample images from the Retinal Wave Dataset (Cappell, 2023); licensed under CC BY 4.0), images have been resized for visualization (A). Representational geometry was computed for sharp ($\sigma = 0$; B) and blurry ($\sigma = 4$; C) versions of the test stimuli for each model. For each model, representational geometry was compared with that of V1 using Spearman correlation, for 1000 bootstrap resamples per cohort. As for the tuning maps, plotted values correspond to the bootstrap median value with error bars indicating the 95% confidence interval. Across both sharp and blurry test stimuli, the retinal wave-trained model was found to be a poorer fit to V1 than the object-trained model, indicating that by 2 months V1 already exhibits tuning to object-specific visual features. Corresponding results for all ventral visual stream regions and test-blur levels, alongside the corresponding model layer are shown in Appendix 3, Figure A3.15.

3.4 Discussion

To characterise the features comprising developing object representations in infancy, we took a NeuroAI approach, combining awake infant fMRI with computational modelling using deep neural networks (DNNs). By modifying the visual experience of DNNs during training, we can probe the types of features that contribute to developing neural responses. By systematically blurring both the training data and the objects used to characterise visual representations, we developed a novel framework for probing the spatial scale of object features encoded by the ventral visual stream (VVS). The resulting multiscale tuning maps were calculated for 2-month-old and 9-month-old infants, and adults, allowing us to characterise the emergence of such tuning across infancy.

The pattern of tuning in the primary visual cortex (V1) was found to be adult-like by 2 months of age, whereas tuning in subsequent visual regions showed protracted development through at least 9 months. Across development, models trained on blurry images and tested on sharp images resulted in the highest correspondence with V1, indicating that the region processes features across a range of spatial scales. Responses were best modelled by intermediate model layers. We do not take this to mean that V1 encodes a single, homogeneous intermediate-level feature set. Rather, these layers likely contain heterogeneous units sensitive to features across a range of complexities (Bau *et al.*, 2020), which together provide a good basis for capturing the representational geometry of V1. The fact that infant V1 already aligns most closely with these intermediate stages of a trained network is striking, given the limited visual experience of infants at this age.

What could account for such complexity of response? A candidate mechanism could be contributions from recurrent processing (Kietzmann *et al.*, 2019). At this young age, however, recurrent connections in V1 are both structurally and functionally immature (Burkhalter, 1993; Siu and Murphy, 2018; Nakashima, Kanazawa and Yamaguchi, 2021). Despite this immaturity, it is possible that these recurrent connections are already contributing to object processing. An alternative possibility is that this complexity is already encoded in the feedforward response. It has recently been shown that, in response to natural images, the feedforward sweep in mature V1 similarly encodes features that best align with intermediate layers of DNNs, indicating the importance of multiple nonlinear processing stages in modelling V1 responses to naturalistic stimuli (Cadena *et al.*, 2019; Miao and Tong, 2024). Neuroimaging modalities with greater temporal resolution, such as electroencephalography (EEG) and magnetoencephalography (MEG), could complement these findings by investigating the temporal dynamics of the complexity of developing responses to disentangle the relative contributions of feedforward and recurrent processing. Due to the limited spatial resolution of these complementary modalities, however, the observed immaturities in extrastriate regions could potentially mask V1-specific patterns.

The pattern of tuning in V1 at 2 months was consistent across term-born controls and preterm-born infants who had spent time in the neonatal intensive care unit (NICU). This suggests that tuning in V1 may be somewhat robust to variation in early visual experience, with neither cumulative postnatal visual experience, which was approximately doubled in the NICU cohort, nor time spent in the atypical visual

environment of the NICU (Marchal *et al.*, 2021), appearing to affect tuning patterns at 2 months. The comparatively high split-half noise ceiling in extrastriate regions does, however, merit further investigation, as it suggests that responses in the NICU group may have been more similar across individuals in these regions. Moreover, the similarity of tuning patterns between NICU and control infants does not preclude an important role for early visual experience in shaping such tuning; for example, group-specific differences may have been present before 2 months. By the time of testing, both groups had already been exposed to the natural statistics of the visual world. Indeed, comparison with a model trained on classes of retinal-wave stimuli indicated that the features encoded in V1 by 2 months already reflect sensitivity to object-specific features beyond the low-level features learned from the visual statistics of retinal waves.

This necessity for object-specific training could suggest that the structure of postnatal visual input over the first two months is shaping V1 responses. This is supported by electrophysiological evidence from kittens showing that V1 receptive fields are sensitive to visual input statistics and are fine-tuned according to the early visual environment (Hirsch and Spinelli, 1970). Similarly in humans, visual experience in the early postnatal months is required to establish fine-scale, retinotopic organization of the early visual cortex, including V1 (Heitmann *et al.*, 2023). Early experience is therefore an established contributor to the refinement of tuning in V1 and as such is a candidate mechanism driving such tuning. Alternatively, the features learned by object-trained DNNs could be capturing innate tuning acquired through evolutionary adaptations, rather than the postnatal experience of the infant, that are not accounted for by the retinal-wave-trained model (Zador, 2019; Kanwisher, Khosla and Dobs, 2023).

In contrast, extrastriate visual regions (V2–V4) and later object-selective regions (VO1, FG4) had not reached adult-like multiscale tuning by 9 months, indicating protracted, hierarchical maturation of multiscale tuning. This pattern is consistent with anatomical evidence that higher visual regions undergo extended structural and functional maturation (Condé, Lund and Lewis, 1996; Reh et al., 2020; Arcaro et al., 2017; Kiorpes, 2015). It also aligns with previous findings that later VVS regions in infants contain category structure as early as 2 months, which continues to refine with development (O'Doherty *et al.*, 2026; Kosakowski *et al.*, 2022), while suggesting that the spatial scale of the features driving these responses is still maturing.

Indeed, tuning patterns in extrastriate visual regions do share some structure with adults by 2 months (Figure 3.2). Interestingly, there were subtle trends that suggested that extrastriate visual regions in infants might be biased toward encoding coarser visual features than adults. While further investigation is required, EEG recordings from 6–8-month-olds viewing faces, toys, scenes and bodies have similarly demonstrated preferential low-spatial-frequency tuning (Xie *et al.*, 2022). Having initially coarser visual representations has been proposed to be adaptive (Turkewitz and Kenny, 1982; Vogelsang *et al.*, 2018), with coarse-to-fine visual experience hypothesised to serve as a functional inductive bias, enabling infants to stabilise global shape representations prior to the maturation of high spatial frequency sensitivity (French *et al.*, 2002; Jinsi, Henderson and Tarr, 2023; Lu *et al.*, 2025b). It is therefore intriguing that such an effect might persist in the features encoded by the EVC beyond the perceptual limits imposed by acuity.

What mechanism might drive the hierarchical maturation of object feature tuning along the VVS? One possibility is that it could reflect a sequence of region-specific critical periods during which cortical circuits exhibit heightened plasticity. In non-human primates, these periods progress sequentially along the VVS (Condé, Lund and Lewis, 1996), suggesting a mechanism through which stable representations in V1 could scaffold later refinement of subsequent regions (Reh *et al.*, 2020).

While the current work leveraged deep neural networks as a tractable means to probe representational structure and investigate feature encoding across spatial scales, our findings also provide empirical grounding for biologically inspired modifications which can be incorporated into future models of the developing VVS. The discovery that multiscale object feature tuning emerges hierarchically relative to V1 calls into question whether the common practice of modelling infant vision as a parallel optimisation process is optimal (Vogelsang *et al.*, 2018; Lu *et al.*, 2025b). Indeed, recent computational evidence suggests that it is unnecessary (St-Yves *et al.*, 2023). Better understanding the relative development of object tuning along the VVS could therefore inform inductive biases, such as layer-specific learning rate schedulers mimicking regional plasticity. Intriguingly, implementing sequential critical periods in DNNs has been found to lead to more stable representation learning (Delrocq *et al.*, 2024), providing promise for such approaches.

An exciting future direction will be integrating these regional learning dynamics with models trained on realistic infant visual input. Infant visual experience is distinct from

that of adults, along at least two broad input dimensions - perceptual fidelity and semantic content – the developmental trajectories of which have both been proposed to create curricula optimised for infant learning (Turkewitz and Kenny, 1982; French *et al.*, 2002; Vogelsang *et al.*, 2018; Vogelsang *et al.*, 2024b; Smith *et al.*, 2018; Cusack, Ranzato and Charvet, 2024; Zaadnoordijk, Besold and Cusack, 2022). The enhanced model–brain correspondence observed in response to blurry stimuli when DNNs were pretrained on retinal waves suggests that structured spontaneous activity in utero may serve as an early stage of such a perceptual curriculum (Cappell *et al.*, 2024). Future work should investigate the specific features learned by retinal-wave-trained and retinal-wave-pretrained models to better understand how such pretraining impacts subsequent object tuning, resulting in enhanced correspondence with the brain. One candidate mechanism would be through instantiating retinotopy. An interesting future direction would therefore be to compare developing visual responses to topographic DNN models of the VVS to see whether they show a similar, or indeed a distinct benefit, for characterising developing object responses, particularly in response to blurry test stimuli (Doshi and Konkle, 2023; Margalit *et al.*, 2024; Lu *et al.*, 2025a). Another interesting possibility would be to explicitly model the trajectory of early visual experience from prenatal spontaneous activity through perceptually degraded early visual experience. The benefit of retinal wave pretraining suggests that other types of prenatal experience could also play a role, such as structured light exposure in utero (Reid *et al.*, 2017).

There are several important considerations and limitations when interpreting the results presented here. First, to focus on how visual features encoded by infant V1 are propagated along the VVS, the stimuli used in this study were designed to ensure

visibility by 2 months. High spatial frequency content was therefore excluded on the basis of acuity estimates for this age group, with the presented stimuli containing a maximum frequency of 5.7 cycles per degree. This ensured perceptibility but precluded inferences about tuning to finer spatial details that are known not to be perceived until later in development. Second, as discussed in Chapter 2, regions of interest were defined at the group level using adult atlas-based parcellations warped into infant space, limiting the certainty with which the results can be attributed to specific ventral stream regions. To achieve stable group-level patterns, the tuning maps presented in this work were derived by averaging across participants. This had the advantage of reducing sources of noise, including imprecise regional definition and residual motion, but it also masked individual variability in both the features encoded and the timing of development. As representational geometry was calculated across all voxels within a region, potential within-region heterogeneity was similarly obscured. Third, this work used the ResNet-50 architecture. Future work could replicate these analyses using other CNNs, such as AlexNet, as well as other families of DNNs, such as vision transformers, to determine the extent to which these findings generalise across model architectures (Kanwisher, Khosla and Dobs, 2023). Related work suggests that the overall pattern of findings is likely to generalise across architectures (Jang and Tong, 2024; Lu et al., 2025b). Finally, because visual regions beyond V1 undergo prolonged maturation, longitudinal imaging across infancy and childhood will be essential for charting their developmental trajectories and for establishing benchmarks that can constrain both biological and computational models of vision.

The tuning maps developed here provide a novel framework for probing the features encoded by the VVS across a range of spatial scales. Using this approach, we demonstrate that multiscale object tuning in the visual cortex develops hierarchically. The maturity and richness of V1 responses by two months challenge the prevailing view that infant vision is limited to low-level visual features. Rather, at this young age, the earliest stage of cortical processing already represents object-specific features across a range of spatial scales, revealing that the representations available to the infant brain are richer than previously appreciated.

3.5 Methods

3.5.1 Participants

This study was conducted as part of the Foundations of Cognition (FOUNDCOG) project, a longitudinal neurodevelopment study which included an MRI scan at two timepoints over the first postnatal year, at approximately 2 and 9 months corrected gestational age (CGA), referred to as the 2-month and 9-month timepoints respectively. Ethical approval for this study was granted by the Trinity College Dublin School of Psychology Research Ethics Committee (SPREC012020-10, approved 25 February 2020), and the ethics committees of the two collaborating maternity hospitals: the Research Ethics Committee of The Rotunda Hospital (REC-2019-018, approved 31 May 2019) and the Research Ethics Committee in the Coombe Women and Infants University Hospital (Study No. 8 – 2019, approved 8 May 2020). Participants were recruited from The Coombe Hospital and the Rotunda Hospital (see Appendix 3, Supplementary Methods section A3.1.1) into one of two cohorts: the control group or the neonatal intensive care unit (NICU) group. To be

recruited into either group, infants were without known congenital abnormalities, visual impairments, or contraindications to MRI. Infants in the control group were born at term (38-42 weeks of gestation) and were without diagnosis of brain injury. Infants in the NICU group were preterm-born infants that had spent some time in the NICU. NICU infants were screened for retinopathy of prematurity (ROP) and were excluded from the study if they had ROP that reached Stage 3 or required laser or injection treatment. If infants were eligible to participate, and no contraindications to MRI were identified, they were invited to attend a scanning session with their caregiver when they were approximately 2 months CGA. At the session, informed consent was obtained from a caregiver on behalf of the infant and further MRI safety screening was conducted.

Unless otherwise stated, control and NICU cohorts were combined for analysis, giving a single cohort per age group. 134 infants (29 NICU) were recruited into the FOUNDCOG Project and attended the 2-month scan with their caregivers. Awake functional MRI (fMRI) data was successfully acquired from 131 infants, who were subsequently invited to return at 9 months. 104 infants returned (79%) with 27 (21%) lost to follow up. Awake fMRI data was successfully acquired from 66 infants at 9 months (63%). Two visual tasks were conducted during awake scanning, selected to maximise infant engagement: a videos task (Vanderwal *et al.*, 2015) which comprised a series of short movies and a pictures task which comprised a series of images of objects. The analyses presented in this Chapter used responses acquired during the pictures task. As the videos task was always presented first, a subset of 113 2-month-olds (median CGA = 2.33 months, IQR = [2.13, 2.67]; 44 female) and 51 9-month-olds (median CGA = 9.2 months, IQR = [8.95, 9.45]; 23 female) participated in the pictures task. Following preprocessing and motion

thresholding, a final sample of 101 2-month-olds (median CGA = 2.33 months, IQR = [2.13 - 2.63]; 37 female) and 44 9-month-olds (median CGA = 9.2 months, IQR = [8.92 - 9.51]; 22 female), in addition to 17 adults, provided data that was used to calculate group-level visual object representations used in this study, and related work (O'Doherty *et al.*, 2026).

3.5.2 Awake infant fMRI for visual neuroscience

Inspired by the technical setup of (Deen *et al.*, 2017; Ellis *et al.*, 2020), visual stimuli were projected directly onto the bore of the scanner, above the eyes of the infant, using an Optoma UHD30 4K Ultra HD projector, located in the tech room behind the scanner. Stimuli were projected through an angled waveguide onto an angled mirror located behind the head coil which projected onto the bore. Prior to projection, stimuli were pre-distorted to minimise the appearance of perceptual distortions in the projected image, caused by the angled projection and curvature of the bore. The videos task was always conducted first as it captured attention quickly, which was important at the beginning of the session. The pictures task was presented second, following which, if an infant continued to engage with the session, the tasks were repeated. Both paradigms lasted approximately 5 minutes. Analyses in the current work are based on group-level responses from the pictures task. Infants participated in a median of 1.5 picture runs at 2 months and 1 picture run at 9 months (see Chapter 2 for further details).

3.5.3 Recording object responses with fMRI

Whole-brain fMRI responses were recorded while participants viewed 36 images of objects (Figure 3.1A). For details of the scanning setup and acquisition parameters, see Appendix 3, Extended Methods, section A3.1.2.

3.5.3.A Paradigm

Objects comprised 3 unique exemplars from 12 categories spanning animate, inanimate large, and inanimate small classifications. The FOUNDCOG experimental paradigm is available at https://github.com/ClionaOD/foundcog_paradigm. Objects were presented in a pseudo-random order, ensuring that a single exemplar from each category was presented prior to the next exemplar of any category. Objects were presented for 3 s, interleaved with a fixation cross which was presented for a jittered duration (0.5–1.5 s). Each run of the paradigm consisted of two repetitions of the objects. Prior to each repetition, an attention getter was used to attract the infants' attention to the display. Objects were segmented and displayed on a background that was perceptually transparent. This was implemented as a black background in the stimulus delivery program, as the absence of light cannot be projected. The absence of a well-defined background encouraged the infant to look at the object, preventing confounds due to the ability to segment objects from background which continues to develop over the first postnatal year (Braddick and Atkinson, 2011). Objects loomed towards the infant to hold their attention, initially appearing in the centre of the display at half of their maximum height and increasing monotonically in size according to a logarithmic function. The maximum object size was defined as the size at which either its width or

height matched the corresponding dimension of the display (width of 29 cm (w); height of 16.3 cm (h). The resolution of the displayed images was 640 × 360 px, and the display was located approximately 27 cm from the eyes of the infant, the maximum spatial frequency in the presented images was therefore 5.7 cpd. A visualisation of how a sample FOUNDCOG object may be perceived by a 2-month-old infant is shown in Appendix 3, Figure A3.1B.

3.5.4 Deep neural network models

3.5.4.A Model training

A standard convolutional neural network architecture, ResNet-50 (He *et al.*, 2015), was trained on the ImageNet-1K dataset (Deng *et al.*, 2009), more than 1.2 million images from 1000 labelled object classes, using supervised learning. Models were trained on visual diets consisting of images blurred with a fixed amount using Gaussian blur. The level of the blur was defined by the sigma (s) of the Gaussian in pixels, which ranged from no blur to high levels of blur (s = 0, 0.5, 1, 1.5, 2, 3, 4, 6 px). Models were trained for 60 epochs using a standard hyperparameter configuration and data augmentation procedure (see Appendix 3, Supplementary Methods section A3.1.5 for details). Where applicable, blur was applied in the PyTorch DataLoader using the torchvision GaussianBlur transform, following a random image crop to 224 by 224 pixels. The default kernel size (k = 9) was used for small values of sigma (s = 0.5, 1, 1.5 px). For larger values of sigma, the kernel size was increased to accommodate the increased filter size, preserving image smoothness (for s = 2, 3, 4, 6 px; k = 13, 17, 21, 31 respectively). Two models were trained for each level of blur to control for idiosyncratic variability in the

representations learned by individual networks, due to factors such as random weight initialization (Mehrer *et al.*, 2020).

3.5.4.B Characterising the spatial frequency content of the visual features extracted by blur-trained DNNs

To quantify the spatial frequency of the visual features extracted by DNNs trained to categorise objects across varying levels of blur, for each level of blur we identified a set of 2048 patches from images of objects (see Appendix 3, Supplementary Methods section A3.1.6; Görgün, Schiele and Fischer, 2025), that evoked high levels of activation in individual neurons in an intermediate model layer. The average power spectra of the preferred patches for each blur level were calculated and compared to a baseline of random image patches (n = 8000).

3.5.5 Representational similarity analysis

Representational similarity analysis (RSA; Kriegeskorte, Mur and Bandettini, 2008) was used to quantify the correspondence between visual representations of the ventral visual stream (VVS) and the models, by calculating the pairwise similarity structure (representational geometry) in response to the 36 objects (using a Pearson correlation) and comparing model and brain responses (using a Spearman correlation).

3.5.5.A Regions of interest

Regions of interest (ROIs) were defined using cytoarchitectonic parcellations from the Jülich atlas, which have previously been cross-validated against functional regions of the

ventral visual stream in adults. ROIs were normalised to age-appropriate templates for each cohort and are visualised in Appendix 3, Figure A3.2. The selected ROIs comprised early visual regions: V1 (hOc1), V2dv (hOc2; referred to as V2 throughout), and V3v (hOc3v; referred to as V3 throughout); intermediate regions: hV4 (hOc4v; referred to as V4 throughout) and VO1 (FG1; Rosenke *et al.*, 2018); and a higher-level region in the fusiform gyrus, FG4. Located further anteriorly along the fusiform gyrus, FG4 spans part of higher-level object-selective cortex that has been reported to overlap with functional regions such as FFA-2, FBA, and VWFA-2, which are associated with faces, bodies, and words, respectively, in adults (Weiner et al., 2017). For all ROIs, voxels from the left and right hemispheres were combined for analysis, as homotopic regions have previously been shown to exhibit similar patterns of visually evoked functional activity in young infants (Ellis *et al.*, 2025).

3.5.5.B Group-level representational geometry

Shared (group-level) visual representations were characterised for each age group in each ROI. fMRI data was pre-processed using a custom pipeline (for details see Appendix 3, Extended Methods section A3.1.3; code is openly available at https://github.com/ClionaOD/foundcog_analysis), including rigid body motion correction, normalisation to an age-appropriate template using an affine transform and motion thresholding. Volumes with a framewise displacement (FD) greater than 1.5 mm were excluded from analysis, as were runs in which more than half of volumes exceeded this threshold. Overall, 15% of 2-month-old runs and 8% of 9-month-old runs were excluded on the basis of motion. A further 14% of runs at both ages were excluded

because they were deemed poor quality by an attending researcher during the scanning session. A final sample of 101 2-month-olds (169 runs), 44 9-month-olds (64 runs), and 17 adults (51 runs) contributed to the group-level visual object representations. For each ROI, a run-wise parameter estimate (beta) was calculated for each voxel for each object, yielding multivariate response patterns for subsequent RSA. Voxels deemed to have unstable parameter estimates were excluded from analysis. Within an age group and ROI, the pairwise Pearson correlations of the multivariate object responses were calculated for each unique subject-run pair, giving pairwise representational similarity matrices (RSMs). The pairwise RSMs were then averaged and symmetry enforced to give a group-level RSM, which characterised the shared representational geometry for each cohort in each ROI.

3.5.5.C Characterising visual representations from DNN activations

To compute representational geometry for the DNNs, the objects at their maximum loom size (640 × 360 px) were placed on a neutral grey background (the midpoint of the 8-bit intensity range) to provide a consistent, neutral baseline across stimuli, and were padded vertically with the same grey so that the object was centred in a square image (640 × 640 px). Images were then resized to 224 × 224 px for passing to the model. Layer-wise representational geometry was calculated for each blurry-trained DNN model (train blur) by extracting activations in response to each of the 36 FOUNDCOG object images (test stimuli) following each ReLU after a convolutional layer (49 layers for the ResNet-50 architecture) and performing a Pearson correlation across the multivariate object

responses. This was repeated for the test stimuli across the same range of Gaussian blurs (σ = 0, 0.5, 1, 1.5, 2, 3, 4, 6 px).

3.5.6 Characterising object feature tuning across spatial scales

3.5.6.A Multiscale tuning maps

A bootstrap resampling procedure was used to compare model and brain responses. Group-level representational geometry was calculated for each age group and ROI across 1000 bootstrap resamples of the data (subject-run pairs sampled with replacement).

Model–brain correspondence was calculated for a given age group and ROI for each combination of train and test blur. As two separate models were trained for each value of train blur, 500 of the bootstrap samples were compared to each train model. Group-level representational geometry was compared to each layer of each model for each value of test blur. The layer with the maximum correlation value was chosen as the best model for the given comparison. Values were then concatenated across equivalent train blur models, giving 1000 values for each train-test combination. For each cohort and ROI, multiscale tuning maps were constructed by taking the median correlation value from this bootstrapped distribution of the maximum layer values for each train-test blur combination. This data-driven approach enabled robust characterisation of tuning to object features across a range of spatial scales and input resolutions, with freely varying layers capturing the appropriate level of visual feature complexity for a given train-test combination.

3.5.6.B Developmental change in model–brain correspondence

To visualise developmental change, difference maps were constructed by subtracting the tuning map of a younger age group from that of an older age group. To test whether the distributions were overlapping, the 95% confidence interval of the difference between each of the 1000 bootstrapped maximum correlation values for each train-test combination was determined. Where this interval did not contain 0, a comparison was deemed significant. Significance was denoted with an asterisk in the corresponding cell of the difference map.

3.5.6.C Associated feature complexity: corresponding DNN layers

For each cell in a given tuning map, the number of the layer (ranging from 0–49) that resulted in the maximum correlation was recorded for each of the 1000 bootstraps. For each train-test combination, a distribution of the best fitting layers could therefore be derived. The best fitting layer for a given train-test combination was estimated by taking the median layer number across the 1000. This median value was plotted to produce a layer map associated with each tuning map (Appendix 3, Figure A3.5A).

As the median layer value varied across train-test combinations within a given tuning map, to derive a single layer value to represent a given tuning map and therefore summarise the complexity of the response (early layer being low complexity and later layer being high complexity) a weighted average was used. Rather than relying on the median maximum layer correlation value and the median layer value, for each of the

1000 bootstraps, both the tuning map, (C; consisting of the correlation values of the highest corresponding layer), and layer map (L; consisting of the layer number of the highest corresponding layer) were derived. The weighted average layer for each bootstrap (wal_b) was then calculated by taking the elementwise product of the correlation values (c) and the equivalent layer number (l) for each combination of train blur (tr) and test blur (te) and normalising by dividing by the sum of the correlation values (Equation 3.1).

$$\frac{\sum_{tr,te} c_{tr,te} l_{tr,te}}{\sum_{tr,te} c_{tr,te}} = wal_b \quad (3.1)$$

This generated a distribution of weighted average layer values for each tuning map, producing a single layer value per tuning map. The median bootstrapped weighted average layer and its 95% confidence interval for each age and ROI is depicted in Figure 3.2F.

3.5.6.D Corresponding model accuracy

Classification accuracy was calculated for each train-test combination of blur by calculating the top 1 classification accuracy of each model in response to the ImageNet validation images, blurred by application of Gaussian blur in the data loader with each of the test blur values (σ = 0, 0.5, 1, 1.5, 2, 3, 4, 6 px). The average accuracy was taken across the pairs of equivalent models (trained with the same value of blur) in response to the same level of test blur, giving a single accuracy value per train-test combination (visualised as a heatmap in Figure A3.8A).

3.5.6.E Scaled covariance between ages

To quantify the similarity in tuning patterns between infants and adults within each ROI, the covariance between the infants and the adults was estimated and normalised by the variance of the equivalent adult region for each of the 1000 bootstraps.

3.5.6.F Testing for motion-related effects

To test whether residual motion could be contributing to the observed developmental differences, analyses were repeated using a subset of runs with a more stringent motion threshold. Runs with a median FD exceeding a conservative threshold of 0.5 mm were excluded from analysis. As before, timepoints in the remaining runs were censored if FD exceeded 1.5 mm. The low-motion group-level RSMs comprised 97 runs from 2-month-olds (57% of the 169 runs included in the primary analysis), with a median FD of 0.23 mm (compared to 0.41 mm in the primary analysis); 51 runs from 9-month-olds (80% of the 64 runs included in the primary analysis), with a median FD of 0.23 mm (compared to 0.29 mm in the primary analysis); and 51 runs from adults (100% of the 51 runs included in the primary analysis), with a median FD of 0.14 mm.

3.5.7 Effect of prematurity on object feature tuning

To test whether the cumulative amount of postnatal visual experience influences the development of object feature tuning in V1 at 2 months, the multiscale tuning maps from infants who were born preterm and spent time in the NICU were compared to those of term-born controls. Twenty-nine NICU infants participated in the FOUNDCOG project, of whom 24 participated in the pictures task at 2 months. After preprocessing, high-quality data remained for 18 infants. One infant was excluded due to their

corrected gestational age (CGA) at scan (4.73 months) being significantly higher than the rest of the cohort, leaving 17 infants for analysis (median CGA = 2.43, IQR = [2.2–2.73]; 8 female). Participating infants were without known structural brain injury and had all experienced a relatively complication-free NICU course, making them well suited for investigating the effects of early visual experience on VVS development. The infants in the NICU group were born on average 8 weeks preterm (median 7.71 weeks, range = [4.0–16.14]) and therefore had an additional 8 weeks of postnatal visual experience compared to controls, which was almost double the amount of cumulative postnatal visual experience at the 2 months scan timepoint. Of these 17 NICU infants, 9 completed 1 run of the pictures paradigm, 6 completed 2 runs, and 2 completed 3 runs. As the sample size of the NICU group was smaller than that of the controls, to compare across groups, bootstrap samples of 17 control infants were generated to match the NICU cohort on the number of completed runs (run-matched controls). Run-matched controls were sampled with replacement from a pool of 81 term-born control infants (median CGA = 2.3, IQR = [2.13–2.53]; 28 female). In this group, 41 infants completed 1 run, 28 completed 2 runs, and 12 completed 3 runs.

3.5.8 Correspondence with retinal-wave-trained models

The retinal wave-trained models used in this analysis were obtained from the ReWaRD project (Cappell *et al.*, 2024) via the open-access repository Zenodo (Cappell and Stoll, 2023). Models consisted of a ResNet-50 architecture trained for 90 epochs on simulated retinal waves (rwave-4096_resnet50_epoch90.pth), on ImageNet-1K (ImageNet-scratch-v1_resnet50_epoch90.pth), or pretrained on retinal waves and subsequently fine-tuned on ImageNet-1K (pt_rwave-4096_ft_ImageNet-v1_resnet50_epoch90.pth).

These models were made available under a Creative Commons Attribution 4.0 International license. A ResNet-50 model with randomly initialised weights was used as an untrained comparison. As before, the models were shown the FOUNDCOG object images (test images) and the resulting activations were extracted from all of the ReLU units following convolutional layers (49 layers). Layer-wise representational geometry was calculated for each model and compared to 1000 bootstrap samples for each ROI for each cohort. The layer resulting in the maximum correspondence with the brain was selected for each model (maximum layer value). The bootstrapped median of the maximum layer value, and its bootstrap confidence interval, is depicted in (Figure 3.3). The analysis was repeated with a greyscale (luminance only) version of the FOUNDCOG object images to test whether colour might be driving the difference between the models.

Chapter 4

Learning Dynamics of Visual Representations Across Spatial Scales

Objects contain information at multiple spatial scales, from high spatial frequency (HSF) detail, such as texture, to low spatial frequency (LSF) structure, such as global shape. In the previous chapter, we found that deep neural networks (DNNs) trained on blurry images, which emphasise LSF features, showed the strongest correspondence with visual cortex (V1–V4) when object representations were characterised in response to sharp images. This pattern was consistent across development and was observed in 2- and 9-month-old infants and adults.

Motivated by this finding, we investigated whether HSF- and LSF-based representations converge with brain responses at different rates during learning. We hypothesised that HSF-based representations would converge with visual cortex quickly, with little or no training, whereas LSF-based representations would require more learning and would converge later in training. We also tested whether extrastriate cortex (V2–V4) would be relatively better captured by LSF-based representations in infancy.

To test this, training and test images were decomposed into low-, mid-, and high-frequency bands. Model–brain correspondence was quantified across training, from untrained to progressively trained models, and across network layers. This allowed us to determine which frequency band best matched visual cortex before and after training, how correspondence changed over training for each band, and which layers drove the strongest correspondence within each band.

In untrained models, HSF-derived representations showed the strongest correspondence with visual cortex. With training, HSF remained the best match and was driven primarily by early model layers. In contrast, LSF-derived representations initially corresponded poorly and showed the slowest, but most substantial, improvement with training, with peak correspondence driven by later layers. These dynamics were broadly conserved across V1–V4 and across age groups. However, the relative strength differed by age: infants showed relatively stronger correspondence with LSF-derived representations than adults, whereas adults were relatively better captured by HSF-derived representations.

These findings suggest that DNNs have an architectural bias towards extracting HSF information. At the same time, the relatively stronger LSF correspondence in infant extrastriate cortex, even when V1 appears adult-like, may reflect counteracting inductive biases in the developing brain that reduce premature reliance on high-frequency object features.

4.1 Introduction

In the previous chapter, we found that training deep neural networks (DNNs) on blurry images and subsequently characterising their visual representations (testing) in response to sharp images led to significantly higher correspondence with the early visual cortex for infants and adults compared to both training and testing on either blurry images or sharp images. That training on blurry images, which contain only low spatial frequency (LSF) visual features enhanced alignment with the visual cortex, was consistent with our hypotheses based on prior literature. However, the observation that this was only the case in the presence of standard, sharp test stimuli was a novel finding, highlighting the complementary contribution of high spatial frequency (HSF) features to visual representations in the early visual cortex (EVC).

The propensity of standard object-trained DNNs to strongly emphasise fine-grained image details, or HSF features, for object recognition, often termed texture bias, has been well documented (Geirhos *et al.*, 2018a; Baker *et al.*, 2018; Heinke *et al.*, 2021). Indeed, this phenomenon was central to the rationale for the blurry training regimes used in the previous study. While standard training leads DNNs to rely on fine details (HSF features), blurry training encourages them to learn more shape-based representations (LSF features; Vogelsang *et al.*, 2018).

It was intriguing that the presence of HSFs at test time was beneficial, even in the absence of any prior exposure to these features, and led us to wonder about the representation of object features across spatial scales at different points during training.

Specifically, what is the trajectory of convergence with the visual cortex for visual representations derived from object features at different spatial scales, and how does this change with development?

To investigate such learning dynamics, it was necessary to directly characterise correspondence with brain responses throughout model training. Consistent with our observations in the previous study (Chapter 3), model–brain convergence has been shown to follow a distinct trajectory from model classification accuracy with respect to the cumulative amount of model training. It has been shown that the component of the visual representations learned by DNNs that converges with the visual cortex is learned very early in training, while continued improvements in object classification performance were not found to confer additional benefit for representational alignment (Scholte *et al.*, 2024).

A limitation of the previous chapter is that by only using low-pass stimuli, it was not possible to isolate the effect of training or testing of HSF object features on their own. In addition, the benefit of mid-range features, which are particularly important for human vision (Solomon and Pelli, 1994), could not be disentangled from those of HSF. Further, while the texture bias is a well-documented and widespread phenomenon among convolutional neural network (CNN) architectures, it is merely that – a bias – and far from absolute. Indeed, we know that standard trained DNNs can, and do, also learn to use coarser object features to some degree (Kubilius, Bracci and Op de Beeck, 2016). Therefore, in this study, to isolate the contribution of object information at different spatial scales, we decomposed stimuli using bandpass filtering into distinct spatial-

frequency (SF) bands. This departs from more ecologically motivated stimulus manipulations that model the types of visual stimuli that are experienced early in development, such as Gaussian blur, used in the previous study (Chapter 3). However, band-limited stimuli provide strong experimental control, making them well suited to testing how information at specific spatial scales shapes neural responses.

Bandpass filtering has long been used to investigate mechanisms of visual processing across spatial scales (Sowden and Schyns, 2006). Psychophysics has shown that spatial vision can be described by multiple partially independent SF-selective channels (Campbell and Robson, 1968; De Valois and De Valois, 1980), although the neural implementation of such channels remains incompletely understood. Similar paradigms in infancy indicate early SF-selective processing. Behavioural evidence suggests that infants may have distinct SF-selective channels by around 12 weeks, but not at 6 weeks of age (Banks, Stephens and Hartmann, 1985), while scalp-recorded visually evoked potentials (VEPs) report frequency-specific responses even in infants younger than 2 months (Fiorentini, Pirchio and Spinelli, 1983; Suter et al., 1994), albeit with greater measurement noise at the youngest ages (Braddick and Atkinson, 2011). Together, these results provide evidence for early SF-selective processing, though age estimates vary across methods.

Spatial frequency manipulations have also been used to test which spatial scales support object recognition. Behavioural work suggests a privileged role for mid-range information, particularly when SF is defined relative to object size rather than in absolute retinal units (Solomon and Pelli, 1994). More recent evidence has extended this account,

suggesting that reliance on a relatively narrow mid-range SF band may generalise across recognition behaviour for many classes of objects (Subramanian et al., 2023). Consistent with these behavioural findings, neural responses and object representations have been found to use information across spatial scales in both early and higher-level visual cortex (Vaziri-Pashkam, Taylor and Xu, 2019; Farhang et al., 2025).

In the current work, participants viewed standard object images containing a range of spatial frequencies, allowing us to probe the features that contribute to object representations while infants viewed naturalistic objects as they would in daily life. By contrast, models were trained on bandpass-filtered images, and their representations were characterised in response to bandpass-filtered versions of the same objects viewed by the infants. This design allowed us to systematically probe the spatial scale of developing visual representations in a controlled manner, and to relate model performance and representational changes to the spatial frequency content available during training and testing.

We hypothesised that in an untrained DNN, visual representations derived from images containing only HSF features would be more similar to brain representations than those containing only LSF features. We further hypothesised that LSF features would take longer to learn, which would be expressed as a more gradual increase in correspondence with the brain for models trained on LSF features relative to those trained on HSF features. As LSFs require larger receptive fields to integrate features across space, we hypothesised that they would therefore be learned in later layers. We the investigated

to what extent representations derived from learning object features at different spatial scales converge or diverge in their resulting representations.

Although this work focused on early visual cortex, we also conducted a preliminary analysis of later visual regions. We hypothesised that model–brain convergence would unfold over a longer timescale in these regions, and that the object-classification performance of the models would be related to the ability of the models to capture brain responses.

Finally, in the previous study, while tuning across spatial scales in V1 was qualitatively mature as young as 2 months, tuning across spatial scales in the extrastriate cortex (V2–V4) remained immature through 9 months. Although the pattern of immaturity was inconclusive, there was suggestive evidence that infants might have coarser representations than adults, even in older infants (9 months), who could perceive all of the spatial frequencies present in the stimuli. Using these delineated bands, we therefore tested the prediction that extrastriate responses in infants would be less well captured by HSF-derived model representations than those of adults.

4.2 Methods

4.2.1 Decomposing images into frequency bands

4.2.1.A Implementation details for bandpass filters

Images were decomposed into three spatial-frequency bands (low, mid, and high) using Butterworth filtering. Butterworth filters were chosen as they provide a gradual

transition around the frequency cutoff, reducing ringing (Gibbs) image artefacts compared with the hard cutoff of a brick-wall filter. This choice prioritised reducing artefacts and perceptual quality over a strict separation between bands. Band boundaries were defined by two cutoff frequencies in cycles per pixel (cpp; where 0.5 cpp is the Nyquist frequency): 0.055 and 0.15 cpp. These cutoffs were empirically chosen to approximately match the frequency attenuation produced by Gaussian blur with σ = 4 and σ = 1 pixels, respectively (for details of the estimation procedure, see Appendix 4, Supplementary Methods A4.1.1). These values were chosen as in the previous study (Chapter 3), application of Gaussian blur with a value of sigma less than 1 was found to lead to deep neural network models (DNNs) that achieved high classification accuracy but were a poor fit for the early visual cortex. This was therefore chosen as the high pass cutoff, while sigma of 4 was chosen to approximate the lower frequency conditions in Chapter 3.

Filtering was implemented using the *skimage.filters butterworth* function (van der Walt *et al.*, 2014) with a squared Butterworth filter (order = 3) and 40 pixels of padding to reduce boundary artefacts. The low band was obtained by low-pass filtering at 0.055 cpp, the high band by high-pass filtering at 0.15 cpp, and the mid band by low-pass filtering at 0.15 cpp followed by high-pass filtering at 0.055 cpp. Images were loaded as 8-bit unsigned integers (0–255). As mid- and high-pass filtering yields values centred on zero, each filtered image was scaled by its maximum absolute value to constrain values to the range [−1, 1], and then mapped back to [0, 255]. This mapping places zero at 127.5 (the midpoint of the 8-bit range), which appears as mid-grey. The low-band images did not require this re-centring. Values were then rounded, cast to 8-bit unsigned integers,

and saved. Centring on mid-grey means that areas with little filtered content appear as a grey background, while lighter or darker regions indicate departures from this baseline. Using a fixed mid-grey baseline also kept the background level consistent across stimuli, helping ensure that comparisons were driven by filtered image content rather than differences in overall brightness.

4.2.1.B Equivalence between the chosen spatial frequency bands and perceptual channels

To relate the image bands to human spatial frequency channels, we estimated the approximate cutoffs in cycles per degree of visual angle (cpd). Bandpass filtering was applied in the DNN input space (224 × 224 px). Using the cutoff frequency ratios of 0.055 and 0.15 cycles per pixel, the corresponding spatial frequencies were 0.055 × 224 = 12.32 cycles per image (cpi) and 0.15 × 224 = 33.6 cpi. When displayed to participants, the stimuli subtended 56° of visual angle (dva) in width. Dividing cycles per image by degrees yields cutoff frequencies of approximately 12.32/56 = 0.22 cpd and 33.6/56 = 0.60 cpd.

As the DNN inputs were down sampled from the 640-px-wide images shown to participants to 224 px (a factor of 2.86), the images seen by the DNN contained less fine spatial detail than those seen by participants, particularly at higher spatial frequencies. However, while the high-frequency band contains less information at the very finest spatial scales due to down sampling, this resizing step does not change the cpd mapping of the selected band boundaries, since these are defined in cycles per image.

4.2.2 Band-specific deep neural network models

A standard convolutional neural network architecture, ResNet-50 (He *et al.*, 2015), was trained on the ImageNet-1K dataset (Deng *et al.*, 2009) using supervised learning. Models were trained on visual diets consisting of images from a single frequency band (low, mid, or high). For comparison, a model was also trained on standard images, henceforth referred to as the all-band model. As decomposing images into frequency bands is computationally intensive, the ImageNet training and validation images were prefiltered into the three bands prior to training using the Butterworth filters defined above.

For each of the four band conditions, an initial network was trained for 60 epochs with a batch size of 256 using a standard hyperparameter configuration. Training used stochastic gradient descent with momentum (0.9) and an initial learning rate of 0.1, decayed by a factor of 10 every 30 epochs. During training, standard ImageNet data augmentation was applied and images were randomly resized and cropped to 224 × 224 px. ImageNet images vary in size, with a median width and height of approximately 500 px and 375 px respectively. As the filtering was applied prior to the crop and the crop scale varies across samples, this step introduces some variability in the effective spatial-frequency content of the final network inputs for the bandpass filtered conditions, however, prefiltering provided a practical way to impose the intended frequency constraints at scale. This approach contrasts with the blurry trained models in Chapter 3, where images were first resized and cropped and subsequently blurred in the data loader. Here, prefiltering was adopted to minimise computational cost while retaining standard ImageNet augmentation.

Images were then randomly flipped horizontally, converted to tensors, and normalised using the standard ImageNet channel-wise mean [0.485, 0.456, 0.406] and standard deviation [0.229, 0.224, 0.225]. Although the per-band channel-wise mean and standard deviation differed from the standard ImageNet values (Table A4.3), we used the standard (all band) normalisation values for all models, to maintain a consistent input scaling and facilitate direct comparability across training conditions.

Training was implemented in Python 3.7.7 with PyTorch 1.7.1 and CUDA 10.1, on a Lambda Quad GPU workstation with 28 CPU cores, 128 GB of system RAM, equipped with 4 NVIDIA RTX 6000 GPUs, each with 24 GB of memory.

4.2.2.A Categorisation accuracy of DNNs trained on different frequency bands

As described above, a single network was trained on each band for 60 epochs. At the end of each epoch, categorisation ability was calculated for each model by calculating its top 1 classification accuracy in response to the ImageNet validation images, filtered into the equivalent frequency band.

4.2.3 Quantifying model–brain correspondence using representational similarity analysis

Representational similarity analysis (RSA; Kriegeskorte, Mur and Bandettini, 2008) was performed (as in Chapter 3, section 3.5.5), to compare visual representations of the

bandpass trained models to the visual representations recorded from the developing ventral visual stream (VVS).

4.2.3.A Participant data

The same group-level visual representations (n = 101 2-month-olds, 44 9-month-olds and 17 adults) were used as in Chapter 3 (see section 3.5 for details regarding participants and data processing).

4.2.3.B Characterising band-specific visual representations from DNN activations

For each band, layer-wise representational geometry was calculated for each DNN model by extracting activations in response to each of the 36 FOUNDCOG object images filtered into the same frequency band that the network had been trained on, following each ReLU after a convolutional layer (49 layers for the ResNet-50 architecture) and performing a Pearson correlation across the multivariate object responses. In this analysis, the train and test band were always matched to isolate the contribution of object features across spatial scales.

4.2.3.C Calculating model–brain correspondence

Model–brain correspondence was defined by the network layer that obtained the highest correlation with the brain region of interest. We focus on the early visual cortex (regions V1–V4, as previously defined in section 3.5.5A). To facilitate comparison across ages, the magnitude of the Spearman correlation was scaled by the appropriate split-half noise ceiling prior to analysis (Table A3.1).

4.2.4 Model–brain correspondence across spatial-frequency bands in untrained DNNs

Untrained ResNet-50 deep neural networks (DNNs) were randomly initialised using the standard *torchvision* implementation (Kaiming/He normal initialisation for convolutional layers). The FOUNDCOG stimuli were decomposed into three frequency bands (low, mid, high) as described above (Figure 4.1A). For each frequency band, we generated 10 independently initialised networks (n = 30 total); each network was presented only with stimuli from its assigned band. For each DNN, brain–model correspondence was quantified as the maximum layer-wise correlation (as defined above).

For each ROI and age group, correspondence was compared across frequency bands using an omnibus Kruskal–Wallis test (3 groups; Table A4.1). Omnibus p-values were corrected across all ROIs and ages tested using the Benjamini–Hochberg false discovery rate (FDR) procedure (12 tests: 4 ROIs × 3 ages). Where the omnibus test indicated an effect of band, post-hoc two-sided Mann–Whitney U tests were performed for all pairwise band comparisons (3 comparisons), with p-values corrected within each ROI and age using the Holm procedure (Table A4.2).

4.2.5 Learning dynamics of object features across spatial scales

4.2.5.A Estimating the approximate trajectory of convergence between model and brain responses

Each epoch consists of a large number of images (1.28 million). Given our batch size of 256 images, each epoch can be subdivided into 5005 batches. To estimate the approximate learning trajectory while capturing early learning dynamics, models were trained for 60 epochs while their representations were characterised on a logarithmic

scale as a function of batch number. Batches were geometrically sampled, with batch number increasing in size by a factor of approximately $\sqrt{2}$. This provided high temporal resolution early in training while still spanning the full training trajectory. To characterise approximate learning dynamics, a logistic function was then fit to this curve.

4.2.5.B Characterising learning dynamics with a logistic fit

For each ROI and age, a four-parameter logistic curve was fit to the trajectory of model–brain correspondence for each band-trained network. Let x denote training time in batches on a logarithmic scale ($\mathrm{x} = \log_{10}(\mathrm{batch})$) and y denote the magnitude of the correspondence with the brain (Equation 4.1):

$$y(x) = y_{min} + \frac{y_{max} - y_{min}}{1 + e^{-k(x-x_o)}} \tag{4.1}$$

Where y_{min} is the lower asymptote, y_{max} is the upper asymptote, k is the steepness and $x_0$is the inflection point (the midpoint of the curve).

Learning dynamics could be subsequently summarised with three interpretable metrics:

1. **Midpoint of Learning**: the inflection point (x_0).
2. **Magnitude of learning**: the change in correspondence between asymptotes ($\Delta y = y_{max} - y_{min}$).
3. **Rate of learning**: the maximum slope of the fitted curve, which occurs at the inflection point:

$$\frac{dy}{dx}\Big|_{x=x0} = \frac{\Delta y\, k}{4} \tag{4.2}$$

4.2.5.B.a Evaluating logistic fit

To assess goodness of fit for each logistic model, we computed the adjusted coefficient of determination (R^2_{adj}). We also compared to a log-linear baseline.

4.2.5.B.b Summarising learning dynamics across bands

Across ROIs and age groups, three metrics of learning dynamics were extracted from the fitted logistic curve (as defined above): the midpoint of learning, magnitude of learning, and rate of learning. When the estimated midpoint lay beyond the training window, midpoint-based estimates were treated as right-censored and were excluded from evaluations of the mid-point of learning.

To summarise general learning dynamics across frequency bands, the distribution of each metric was compared across bands using an omnibus Kruskal–Wallis test (3 groups). Where the omnibus test indicated an effect of band, post-hoc two-sided Mann–Whitney U tests were performed for all pairwise band comparisons (3 comparisons), with p-values corrected using the Holm procedure within each metric.

4.2.5.C Robustly characterising early learning dynamics

To robustly characterise early learning dynamics for each spatial frequency band, 10 independently initialised instances of the networks were trained on visual diets consisting of images from a single frequency band (low, mid, or high) for a short duration of 3 epochs. This was estimated to incorporate the mid-point of learning for each of the bands based on a preliminary analysis (see section 4.2.5.A) As before model

representations were characterised on a logarithmic scale as a function of batch number. In this instance, however, due to the computational load of training 10 instances with 4 band configurations, the models were minimally trained for the first three epochs, a period that we had found in pilot work to contain a substantial portion of the overall learning. We refer to this period as the critical training window. Following which, a logistic function was then fit to each curve to characterise learning dynamics (as in section 4.2.6.B).

4.2.5.D Comparing representational correspondence before and after DNN training

For each combination of ROI, age and frequency band, we performed a within-network paired comparison of correspondence at the untrained timepoint (epoch 0) versus the endpoint of the critical training window (epoch 3) using a Wilcoxon signed-rank test. We then controlled the false discovery rate across the full set of 36 tests (4 ROIs × 3 ages × 3 bands) using the Benjamini–Hochberg procedure ($\alpha = 0.05$).

4.2.5.E DNN layer comparisons

Spearman correlation was used to compare representations across layers from models presented with and trained on stimuli from different bands. Models trained on a given band were always presented with stimuli filtered into the same band when characterising object representations.

4.2.6 Band-dependent differences between infants and adults

Brain–model correspondence was evaluated at the endpoint of the critical training window (epoch 3) following the same procedure as in section 4.2.4. Correlation coefficients were Fisher z-transformed ($z = \operatorname{atanh}(r)$) prior to inference. For each ROI and age group, band differences were first assessed with an omnibus Kruskal–Wallis test across the three frequency bands. Omnibus p-values were corrected across all ROI × age tests using Benjamini–Hochberg FDR. Where significant, follow-up two-sided Mann–Whitney U tests were run for all pairwise band comparisons, with Holm correction applied within each ROI × age.

To quantify developmental differences in frequency-band effects, we compared adults to infants within each ROI (adult–2 month and adult–9 month contrasts). Correlation coefficients were Fisher z-transformed ($z = \operatorname{atanh}(r)$). To remove any region-level band-invariant offset and focus on how the correspondence varies across bands. z-values were demeaned across bands for each network iteration (n = 10). Adult–infant contrasts were then computed, yielding a demeaned Δz for each band and iteration, reflecting deviation from the region's across-band mean adult–infant effect. As we had a small number of iterations (n = 10) support for parametric assumptions was limited, thus we tested whether the mean demeaned Δz differed from zero using a two-sided exact sign-flip permutation test (enumerating all sign configurations). 95% confidence intervals for the mean were estimated using bootstrap resampling across iterations (5,000 resamples). P-values were corrected using Benjamini–Hochberg FDR across ROI × band tests within each adult–infant contrast (12 tests per contrast).

Positive values indicate that the band provided a relatively better fit for adults than infants relative to the ROI's across-band mean, whereas negative values indicate a relatively better fit for infants than adults.

4.3 Results

4.3.1 In untrained deep neural networks, high spatial frequency visual features best capture early visual cortex responses across development

Untrained deep neural networks (DNNs) have not yet learned to extract object-specific features. Comparing visual representations from untrained DNNs in response to images containing different spatial frequencies enables us to investigate how the DNN architecture (ResNet-50) might intrinsically capture components of early visual cortex representations in response to visual information at particular frequencies. We hypothesised that visual object representations extracted from untrained DNNs in response high spatial frequency information would be more like the visual object representations encoded by the early visual cortex than those extracted in response to low spatial frequency information.

To compare model representations in response to information across different spatial frequencies, the FOUNDCOG test stimuli were decomposed into three spatial frequency bands containing low, mid and high spatial frequency (LSF, MSF and HSF) information (Figure 4.1A). Untrained DNNs were randomly initialised, and each was given as input images from a given frequency band, such that layer-wise visual representations could

be calculated in response to visual information across spatial frequencies. Each stimulus set was presented to 10 independent random initialisations of untrained DNNs. Across development (2 months, 9 months and adults), the magnitude of correspondence between visual representations in the early visual cortex (V1–V4) and representations extracted from the DNN, was found to be significantly different depending on the spatial frequency of the features presented to the model (Figure 4.1B; Table A4.1). Post hoc tests revealed that representations derived from HSF object features were significantly better at capturing early visual responses than MSF object features, which were in turn significantly greater than LSF features. This was the case for all comparisons, with the exception of infant V4 where representations derived from MSF and LSF information were not significantly different from each other (Table A4.2).

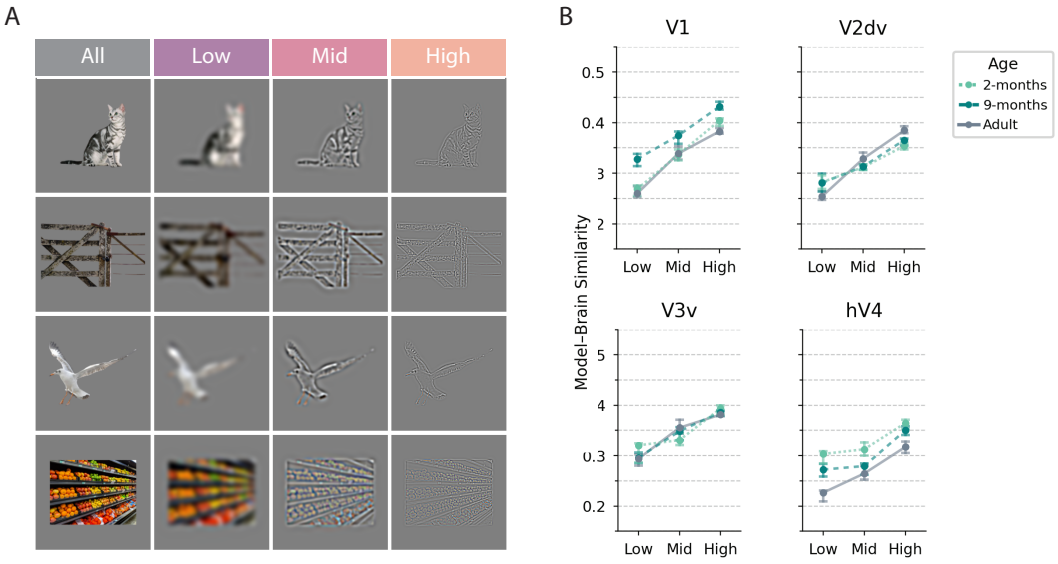

Figure 4.1 In untrained deep neural networks, high spatial frequency object features best capture early visual cortex responses across development

(A) Sample FOUNDCOG stimuli decomposed into frequency bands containing low, mid, and high spatial frequency information. (B) Stimuli from each band were given as input to untrained deep neural network models (n = 10 random

initialisations per band). For each band, model–brain similarity (correlation) with early visual cortex responses was calculated separately for each layer, and the maximum layer correlation was selected. Markers show the median maximum layer correlation across model initialisations, and error bars show 95% confidence intervals. In these untrained models (B), visual representations derived from high spatial frequency features captured early visual cortex responses significantly better than those derived from mid spatial frequency features, which in turn performed significantly better than those derived from low spatial frequency features.

4.3.2 Learning trajectories of object features across spatial scales

4.3.2.A Spatial frequency information across spatial scales contains learnable object features

As for the FOUNDCOG stimuli, training images for the DNNs were decomposed into three frequency bands (sample images are visualised in Appendix 4, Figure A4.1A). DNNs were trained on images of a given spatial frequency, creating models that learned object-relevant features at a given frequency band (band-specific models). To verify that information in each band contained learnable object features, networks were trained for 60 epochs, and the accuracy of each model was determined throughout training. Model accuracy was found to differ across bands but was reasonable for each, with all bands obtaining over 60% top 1 accuracy following 60 epochs of training. All bands therefore contain learnable category-relevant object features (Appendix 4, Figure A4.1B).

4.3.2.B Preliminary band-wise trajectory of convergence with early visual cortex

We next sought to investigate the band-wise learning dynamics of the networks with respect to their trajectory of convergence with responses measured from the early visual cortex (EVC). We hypothesised that low spatial frequency (LSF) features would be learned in intermediate to late layers, would take longer to learn, and would therefore converge relatively slowly with the EVC. While high spatial frequency (HSF) features, which we have shown above correspond better even in untrained DNNs, would be learned in early layers and would converge more quickly with EVC responses.

To investigate the trajectory of convergence between the band-specific models and the EVC, it is necessary to directly measure the correspondence between them. As we have previously shown (Chapter 3), the accuracy achieved by a given model does not necessarily predict its correspondence with the brain.

DNNs are typically trained on a large number of epochs, each consisting of a large number of images. The motivation for this is typically to obtain the highest possible accuracy on an object classification task. However, it has been shown that convergence between model and brain responses can occur much earlier in training (Scholte *et al.*, 2024). Indeed, a substantial amount of learning is observed in the earliest epochs (Appendix 4, Figure A4.1B). It is therefore possible that convergence with the early visual cortex might occur rapidly in the earliest training epochs. Sampling model–brain correspondence early in the training trajectory could therefore potentially be important to capture band-wise learning dynamics.

Each epoch of training consists of more than 1 million images but can be subdivided into smaller batches of images. To estimate the trajectory of convergence with the brain while capturing early learning dynamics, we therefore characterised model representations on a logarithmic scale, as a function of batch number, providing high temporal resolution sampling frequency early in training. This allowed us to perform an initial sweep of the full training trajectory to estimate the dynamics of convergence with the EVC. Representations at each stage were then compared to EVC (V1–V4) responses from infants and adults (Appendix 4, Figure A4.1C).

To characterise the approximate learning dynamics, a logistic function was fit to each curve (Appendix 4, Figure A4.1C). From the logistic fit, the following parameters summarising the band-wise learning dynamics could be estimated: the midpoint of learning, the magnitude of learning and the rate of learning (Appendix 4, Figure A4.1D-F). In particular, we sought to determine whether the logistic function was a reasonable fit and, if so, the approximate midpoint of learning to inform the appropriate sampling window for more robustly characterising band-wise learning trajectories. In general, a logistic function was found to be a good fit for the trajectory of correspondence in the low and mid bands (high adjusted R^2 values) and was typically a better fit than a simpler loglinear alternative (Table A4.4). The high band was generally less stable but was reasonably fit by a logistic function in adults. This initial sweep suggested that there were potential interesting differences in the trajectories of convergence across bands (Table A4.5; Table A4.6). Convergence with the EVC was estimated to occur very early in training, during the initial epochs.

4.3.3 Characterising band-wise learning trajectories

4.3.3.A Band-wise learning dynamics

Building on the encouraging preliminary findings above, we sought to more thoroughly characterise band-wise early learning trajectories. We therefore trained several DNNs per band (n = 10) for a duration of 3 epochs, using the same batch-wise logarithmic sampling scheme as above with an untrained starting point. The three completed epochs in the sampling window were also included as these datapoints have uniformly sampled all batches and are therefore likely to be particularly stable datapoints. We chose this training window as piloting (Methods Section 4.2.5.A) indicated that it captured important early learning dynamics, we thus refer to it as the critical training window. Training several models per band enabled us to extrapolate general band-wise trends beyond individual differences in the models (Mehrer *et al.*, 2020). In particular, the earliest stages of training are likely to be unstable, due to the stochastic nature of visual experience of the model, which consists of a partial, random sample of the training data.

Across all frequency bands, model–brain correspondence was found to significantly increase over the critical training window, from the untrained starting point (epoch 0) to the endpoint (epoch 3), for every age group in each ROI (within-network paired Wilcoxon signed-rank tests; Benjamini–Hochberg FDR correction across all 36 ROI × age × band comparisons, $\alpha = 0.05$). Across regions, model–brain correspondence increased the most for DNNs trained on LSFs, followed by those trained on MSFs. Correspondence with DNNs trained on HSFs increased the least over this training period, however, these increases were still found to be significant (Table A4.10).

As above, a logistic function was fit to each curve enabling characterisation of the dynamics of convergence of model–brain responses (Figure 4.2A). A logistic function generally fit the data well. As a baseline, a loglinear fit was also performed. Adjusted R^2 values were calculated for both fits, and the best fitting curve was adjudicated by comparison. Across all ages, correspondence between the EVC and LSF trained models was found to be well fit by a logistic function as indicated by high adjusted R^2 values (Table A4.7**)**. Correspondence with HSF trained models was less well fit by the logistic curve, however, the logistic curve remained a superior fit to the loglinear curve, as adjudicated by comparison of adjusted R^2 values. The MSF curve was also fit quite well by a logistic, particularly for the adults. The logistic function was found to be a poorer fit for the high band but remained superior to the loglinear fit and was therefore a reasonable approximation for adults. As the logistic was found to be a reasonable fit for all conditions, interpretable band-wise learning dynamics could be derived from the fits: the midpoint of learning, the magnitude of learning and the rate of learning.

To evaluate band-specific learning dynamics, band-wise summary metrics derived from the logistic fits were aggregated across region and age (Figure 4.2B-D). The midpoint of the logistic was taken as a proxy for the time of learning. Across all bands, the midpoint of model–brain convergence occurred early in training, within the first epoch. As hypothesised, the low band took the longest to reach its midpoint of convergence (median = 4.29×10^5 images, 95% CI = [4.18×10^5, 4.36×10^5]), while the high band was almost 2 orders of magnitude faster (median = 8.4×10^3 images, 95% CI = [8.09×10^3, 9.05×10^3]). Interestingly, however, the mid band converged even more rapidly (median

= 4.34×10^3 images, 95% CI = [3.6×10^3, 6.88×10^3]) than the high band, despite starting at a significantly lower level of correspondence (Figure 4.2B). The magnitude of learning was also found to be significantly different between bands, with the low band increasing the most with training, the high band increasing the least and the mid band increasing an intermediate amount (Figure 4.2*C*). The rate of learning was not found to differ across bands.

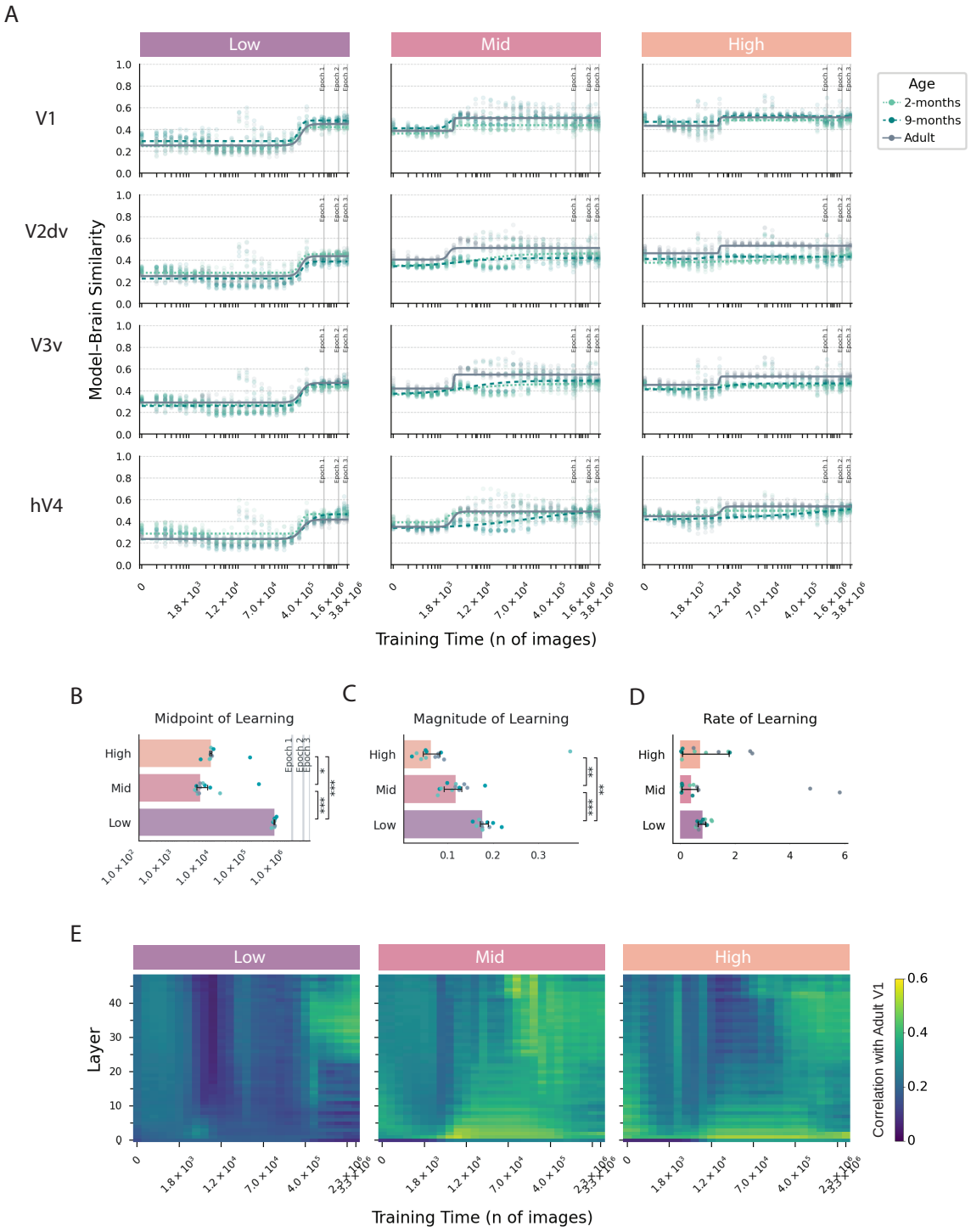

Figure 4.2 Characterising the band-wise trajectory of convergence with the early visual cortex across development

Model–Brain correspondence was determined between the developing EVC (V1–V4) and models trained on object features across spatial scales. Convergence with

the EVC was found to occur early in training. To evaluate the dynamics of convergence, logistic curves were fit to each trajectory (*A*). An equivalent plot to (*A*) of the median and 95% confidence of each curve is depicted in (Appendix 4, Figure A4.2). The logistic fit enabled band-wise learning dynamics to be extrapolated, including the mid-point of learning (B), the magnitude of learning (C) and the rate of learning (D). To visualise which layers were driving responses across the training window, the layer-wise correspondence was visualised. The pattern of layer-wise correspondence was found to be similar across the EVC across development and is therefore depicted for a representative regions – adult V1 (E). All regions (V1–V4) are depicted for each cohort in (Appendix 4, Figure A4.3).

4.3.4 Low spatial frequency object features are learned in later layers, while high spatial frequency features peak in early layers

4.3.4.A Layer-wise trajectories of correspondence with the brain by band

What layers are driving the model–brain correspondence? Do they differ across band or change with cumulative amount of training? We hypothesised that LSFs object features would be learned in later layers while HSFs object features would be learned in earlier model layers.

Model–brain correspondence for each layer across the critical training window was derived for each band-wise model iteration. The average pattern of layer-wise correspondence was qualitatively similar across ages and EVC regions and is visualised for a representative region – adult V1 (Figure 4.2E; Further regions (V1–V4) across development are depicted in Appendix 4, Figure A4.3).

As hypothesised, low spatial frequency features, which were shown above to take the longest time to converge with the EVC, were found to build up in later network layers. High spatial frequency object features were best capture by a peak in very early layers, centred around the second model layer which appeared very early in training. This early perceptual peak was found to lead to the best correspondence with the EVC for high spatial frequency object features over the critical training window. However, correspondence was also found to increase in later layers, which likely reflects the model learning category structure. The mid band showed a similar pattern as the high band consisting of both an early perceptual peak and a later build up likely reflecting category structure.

4.3.4.B Convergence of model representations learned from features at different spatial scales

How similar are the representations derived from object features across spatial scales? Although the DNN models were presented with different object features, it is possible that the derived visual representations are convergent, particularly in later layers following training, as models learn to derive features following the same objective – object classification.

To quantify layer-wise similarity in model representations derived from features at different spatial scales, we performed layer-wise spearman correlation across models for all layers of the untrained networks (Figure 4.3A) and at the endpoint of the critical training window (Figure 4.3B). To evaluate the degree to which representations derived from a given band changed due to training, we also compared layer-wise representations derived from each model before and after training (Figure 4.3C).

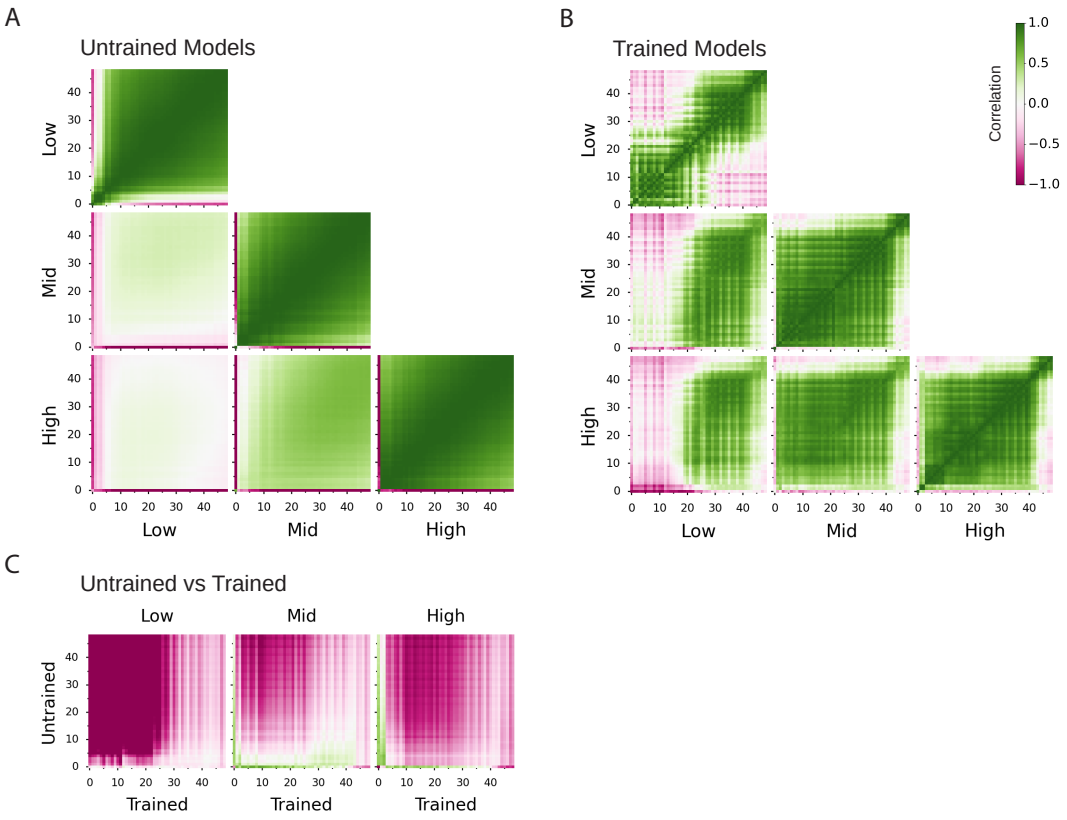

Figure 4.3 Convergence of model representations learned from features at different spatial scales.

Layer-wise representations were extracted from untrained DNNs in response to stimuli containing low, mid or high spatial frequency features (n = 10 random initilizations per band). For a given band, responses were averaged and then compared to layer-wise representations from the other bands (Spearman correlation) to determine the similarity of visual representations derived from each band (A). This procedure was repeated following the critical training window (3 epochs). Visual representations were calculated for the same band on which the model was trained (B). The similarity of representations for a given band before and after training were also compared to evaluate the degree to which layer-wise model representations changed with training (C).

Prior to training, the features encoded by each model were similar across layers. The features encoded by the models presented with LSF object features were quite different from those encoded in response to MSF and HSF object features, which were somewhat more similar to one another (Figure 4.3A). Following the critical training window, the later layers of the model trained on LSF images were similar to all of the layers of the model trained on HSF images, indicating that the layers that best corresponded with the brain had relatively similar representations (Figure 4.3B). In general, model representations were found to be quite different before and after training Figure 4.3C), with one intriguing exception – the early perceptual peak. The peak comprised the first three layers of the model trained on HSF images and led to the highest correspondence with the EVC was found to have quite similar representations to the untrained network presented with HSF object features, indicating that these representations changed little with training (Figure 4.3C).

4.3.5 Band-dependent differences across development

To investigate whether features at a particular spatial scale were a better fit to early visual responses across development, we sought to determine whether, at the endpoint of the critical training window, networks trained on low, mid or high spatial frequency features, and tested on equivalently transformed stimuli, would better fit the EVC across development. In Chapter 3 there were trends suggesting that responses from infant extrastriate regions (V2–V4) might correlate more with representations derived from training on LSFs than those from adults, however, the results were inconclusive. Using the band trained models, we can more specifically test the similarity to representations derived from object features across spatial frequencies. Given the findings from Chapter

3, we hypothesised that V1 would be consistent across ages but that in infants, extrastriate cortex (V2–V4) would be best fit by the DNN that had learned to extract LSF object features. We also hypothesised that correspondence with representations from the HSF network would improve with development.

In adults, correspondence with the EVC was significantly higher for models trained on HSFs than those trained on LSF and MSF features, which were not significantly different from one another (Figure 4.4A; Appendix 4, Tables A4.11 and A4.12). This was also the case in infant V1. Extrastriate regions (V2–V4) had subtle differences for infants, whereby the HSF trained model was less well differentiated from LSF and MSF trained models.

To further investigate these putative developmental differences, controlling for potentially unmodelled factors that could be driving overall differences in the magnitude of correspondence across ages, for each age correspondence was demeaned across all bands. Infants were then subtracted from adults to give a measure of relative fit for 2-month-olds and 9-month-olds relative to adults (Figure 4.4B; Appendix 4, Table A4.13). Across the extrastriate cortex (V2–V4), infants were found to be better modelled by the low band than adults, reaching significance that survived correction in infant V4 (at 2 months and 9 months) and 2 month V2. The inverse was also found to be the case with adults being better relatively modelled by the high band than infants, reaching significance in V2 and V3 relative across infancy and in V4 relative to 2-month-olds.

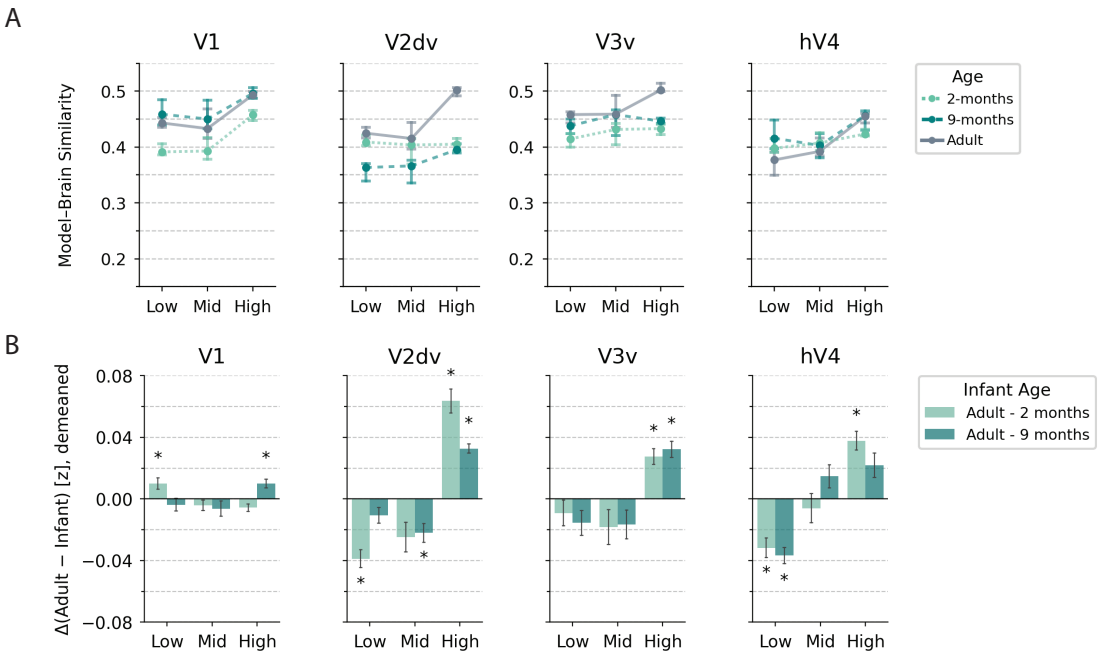

Figure 4.4 Representations learned from low spatial frequency object features were a relatively better fit for infants, while those learned from high spatial frequency object features were a relatively better fit for adults.

(A) Model–brain similarity between early visual cortex responses and model representations at the endpoint of the critical training window was calculated for each spatial frequency band. Across age groups, representations derived from high spatial frequency features showed the strongest model–brain similarity in early visual cortex. (B) To compare model–brain similarity between infants and adults, correlation values were Fisher z-transformed, demeaned across bands within each age group, and then subtracted between age groups. Across extrastriate cortex (V2–V4), infants were relatively better modelled by representations derived from low spatial frequency features, whereas adults were relatively better modelled by representations derived from high spatial frequency features.

4.3.6 Trends in later ventral visual stream regions

As an exploratory analysis, we compared later ventral stream regions, specifically intermediate regions LOC and VO1 and category selective region FG4 to the preliminary models trained on features across spatial scales for 60 epochs (defined in section 4.2.5.A). As these regions encode object shape and category, and as we have previously shown, correspond with later network layers (Chapter 3; O'Doherty *et al.*, 2026) we hypothesised that they would continue to increase with further model training as accuracy increases, particularly for adults as we have shown these regions to be immature in infants (O'Doherty *et al.*, 2026). Preliminary results suggest that there may indeed be continued increases with additional training (Appendix 4, Figure A4.4A-B). Further, a preliminary layer-wise analysis of adult regions revealed that model–brain correspondence in these regions appears to be mediated by later layers and was found to build up later in training (Appendix 4, Figure A4.4C).

4.4 Discussion

Deep neural networks (DNNs) were trained on objects that were decomposed into features across a range of spatial scales. Visual representations were extracted throughout training and were compared to the representations derived from the early visual cortex (EVC) across development. In this way, the trajectory of model–brain correspondence was characterised for models that had learned features at each of these scales, describing the learning dynamics of brain–relevant features across each of these scales.

As hypothesised based on the previous study (Chapter 3), even in untrained networks, visual responses derived from objects containing only high spatial frequency (HSF) information, corresponded more with the EVC than those derived from mid or low spatial frequency (MSF or LSF) information. This correspondence was driven by three layers early in the model, which we therefore termed the early perceptual peak. Over the critical training window, in response to HSF stimuli, this peak continued to dominate representational correspondence with the EVC and was found to remain similar in its representational content to the untrained model (Figure 4.3C).

Visual representations learned from LSF object features took the longest time to converge with responses from the EVC, driven by representations in later layers. Prior computational work investigating how spatial frequency components are learned in DNNs, indicated that LSFs are learned first while HSFs are learned later (Xu, Zhang and Xiao, 2019; Xu, Zhang and Luo, 2025). This seeming contradiction with the rate of learning observed in the present work emphasises the importance of directly investigating convergence with the brain response, when that is the metric of interest. The components of visual representations that drive convergence with the brain are a subset of those learned by the model to perform tasks such as object categorisation. Thus, while object training is a useful heuristic for guiding models to learn the types of features that the VVS learns (Yamins *et al.*, 2014) it can only take us so far. The visual features of relevance to the brain appear to follow distinct learning dynamics and, as we and others have shown, extended training on object classification does not necessarily lead to increased correspondence with the brain (Scholte *et al.*, 2024). Further, the observed convergence with HSFs in this work was echoed in untrained DNNs, indicating

that representational alignment with HSFs was at least party driven by inductive biases inherent in the model architecture, rather than learned object features. Over the investigated critical window, in HSF trained networks, this early perceptual peak continued to lead to the highest model–brain correspondence across the EVC. With further training, however, increases in correspondence with later layers were also observed. It would be interesting to further explore convergence with category regions as, unlike the EVC, they did not correspond with early layers, that is, they did not exhibit an early perceptual peak. Convergence with HSF features therefore appears to follow a bi-phasic pattern of correspondence, whereby object-specific HSF features, which may be particularly relevant for correspondence in later VVS regions are learned later. However, in a recent study of model–brain convergence using standard models trained on all spatial frequencies even the latest layers tested were found to plateau in their correspondence with a higher visual region (the PPA) following just 1 epoch of training (Scholte *et al.*, 2024).

Both training regime and convolutional neural network (CNN) architecture have been implicated in the texture bias observed in standard CNNs (Doerig *et al.*, 2020; Wang *et al.*, 2020; Doshi *et al.*, 2025). Our results support a role for both factors. In particular, the early perceptual peak that we observe is consistent with the possibility that architectural constraints may bias models toward preferential use of high-spatial-frequency (HSF) features. This raises a more fundamental question: even if CNNs are limited in their ability to learn about shape, specifically global shape (Doerig *et al.*, 2020), is this necessarily a limitation for a model intended to capture computations within the ventral visual stream (VVS)?

The shape-bias paradigm (Geirhos *et al.*, 2018a)—in which models are evaluated on cue-conflict images (e.g., a cat shape with an elephant texture) according to whether they classify by shape or texture—has been highly influential in NeuroAI. It has generated valuable insights and provides a clear behavioural signature on which humans and models can diverge. Consequently, achieving a more human-like shape bias has increasingly been considered as complementary to, or even preferable to, maximising object classification accuracy. This has motivated a growing literature exploring developmentally inspired training modifications aimed at increasing the shape bias of DNNs (Yoshihara, Fukiage and Nishida, 2023; Müller *et al.*, 2024; Vogelsang *et al.*, 2025; Lu *et al.*, 2025a). However, the metric also has important limitations, particularly as a criterion for evaluating models of the VVS.

Although the paradigm captures a robust behavioural difference between object-trained CNNs and human observers, behavioural responses to this task are not uniquely attributable to VVS computations. Human object recognition can draw on rapid LSF processing and interactions across an extended network beyond the VVS, including top-down facilitation from frontal regions (Bar *et al.*, 2006). In addition, global shape perception has been argued to depend in part on dorsal-stream contributions and may not be computed solely by feedforward representation in the VVS (Ayzenberg and Behrmann, 2022). Consistent with this broader view, several studies suggest that ventral representations are often better characterised in terms of feature-based information (Long, Yu and Konkle, 2018; Bao *et al.*, 2020; Jagadeesh and Gardner, 2022; Loke *et al.*, 2025), than by explicit encoding of global spatial configuration (Ayzenberg and

Behrmann, 2022). This does not imply that shape is irrelevant to the VVS, rather, it suggests that behavioural shape bias might be better interpreted as a complementary measure of a broader perceptual decision-making system, rather than indexing VVS representations. Taken together, these considerations motivate caution in using shape bias, or any behavioural metric of global-shape performance, as a primary criterion for evaluating models of the VVS. The field may therefore be too quick to dismiss CNN architectures on the basis of their limitations in producing 'human-like' global-shape behaviour.

Interestingly, a novel visual anagram task designed to more robustly characterise configural processing in deep networks found that CNNs, which are biased toward local receptive-field processing, showed weaker sensitivity to global structure, whereas transformers architectures enabling long-range integration via self-attention performed better (Doshi *et al.*, 2025). This pattern is consistent with the possibility that successful performance on such tasks reflects contributions from an extended cortical processing network beyond the VVS rather than ventral-stream computations alone.

More broadly, to the extent that ventral representations emphasise feature-based information and integrate global structure via interactions with other pathways, architectural properties that promote reliance on local features may, to some degree, contribute to a model's ability to predict VVS responses rather than detract from it.

As noted above, a rapidly growing literature has explored developmentally inspired training regimes for DNNs (Vogelsang *et al.*, 2024a; Vogelsang *et al.*, 2018; Avberšek,

Zeman and Op de Beeck, 2021; Jang and Tong, 2021; Jinsi, Henderson and Tarr, 2023; Yoshihara, Fukiage and Nishida, 2023; Vogelsang *et al.*, 2024b; Müller *et al.*, 2024; Vogelsang *et al.*, 2025; Lu *et al.*, 2025b; Vogelsang, Vogelsang and Sinha, 2026). This line of work was motivated in part by experimental evidence from individuals born blind due to dense bilateral congenital cataracts (Vogelsang *et al.*, 2018). Following lens-removal surgery, these individuals rapidly receive comparatively higher-acuity input than typically developing newborns (Maurer *et al.*, 1999). These individuals later show characteristic impairments in aspects of configural shape processing (Maurer, Mondloch and Lewis, 2007). These deficits have been hypothesised to emerge as a consequence of the absence of an initial period of visual experience dominated by LSFs (Vogelsang *et al.*, 2018). The early perceptual peak that we observe in CNNs following high-resolution training is consistent with this account and suggests a possible computational mechanism. Specifically, the early availability of highly informative HSF signals may drive learning toward locally sufficient solutions that are advantageous early in optimisation but potentially suboptimal in the long term, limiting the later emergence of broader receptive fields and spatial integration.

Consistent with this, efforts to train DNNs in more human-like ways, such as blurry pretraining, have been shown to produce modest but reliable improvements on behavioural metrics such as shape bias, which is notable given the simplicity of the intervention and the many unmodelled factors (innate and learned) that shape the human VVS. However, particularly for object recognition, these gains can be somewhat transient, once networks are trained on standard high-resolution inputs, they may partially revert, reflecting a continual-learning problem (Jang and Tong, 2021). One

possibility is that this reversion occurs because, as discussed above, HSF features provide an especially efficient signal for the specified category learning task once they become available.

In addition to characterising trajectories of convergence of visual representations learned from object features across spatial scales, we also sought to investigate whether there were developmental differences in the correspondence with visual representations learned from object features at different spatial scales.

Replicating our findings from the previous study (Chapter 3), V1 was found to be qualitatively mature. Interestingly however, in the extrastriate cortex there were statistically significant relative improvements in fit for the LSF model for infants relative to adults and the HSF model for adults relative to infants. These effects were similar in 2-month-olds and 9-month-olds.

This persistence of these lower SF representations in the infant extrastriate cortex over the first year suggests that factors beyond acuity are at least partially responsible for the sustained LSF enhancement in visual representations. As discussed in the previous study (Chapter 3), one proposed benefit of LSF representations is enhanced learning of categories (Jinsi, Henderson and Tarr, 2023). Having lower SF representations in early visual cortex could facilitate this goal, perhaps, for example, by passing this input to later object selective regions in the VVS, biasing the input they receive as they continue to develop. Future work should investigate whether this effect persists into childhood. Interestingly, the developmental shape bias (which, it should be noted, has distinct

differences from the computational shape bias paradigm; Tartaglini, Vong and Lake, 2022), has been found to emerge in the second postnatal year (Landau, Smith and Jones, 1988; Smith, 2003) indicating protracted development of cortical shape processing. An interesting question for future research would be whether the cortical trajectory of shape processing converges with the behaviour emergence of this phenomenon.

This work has several limitations. For example, participants were only shown stimuli containing all spatial frequencies (broadband stimuli). An fMRI study which showed both broadband, lowpass and highpass filtered objects to adult participants reported stronger average responses in early visual regions (V1–V4) for HSF filtered objects (above 4.40 cpd) as compared to LSF filtered objects (below 0.62 cpd; Vaziri-Pashkam, Taylor and Xu, 2019). This pattern of results was consistent with our finding that regions across the adult EVC better corresponded with visual representations derived from HSF than LSF or MSF trained models. This convergence provides encouraging validation for our computational approach. Although note that the band cutoffs were much lower in our study, with LSF stimuli consisting of frequencies less than 0.22 cpd and high spatial frequency stimuli consisting of frequencies greater than 0.60 cpd. Thus, the LSF condition in (Vaziri-Pashkam, Taylor and Xu, 2019) is comparable to our LSF and MSF conditions combined, while our HSF band contains frequencies between 0.60 cpd and 5.7 cpd, which has a broader lower range and truncated upper range. As with the previous study, the upper range of SFs was limited such that the presented SFs can generally be perceived by 2-month-olds.

A further limitation is that learning dynamics were characterised using only a single convolutional neural network (CNN) architecture. ResNet-50 was selected because it is a widely used CNN that has been extensively employed in studies of visual object recognition. Future work should examine whether the observed findings replicate across CNN architectures and extend to other families of DNNs, such as vision transformers. This would clarify whether the inductive biases identified here reflect CNNs as a model class, object-trained DNNs more broadly, or idiosyncrasies specific to ResNet-50.

Despite its limitations, this work also had distinct strengths highlighting the importance of direct comparison with visual representations when fit to the VVS is the metric of interest and providing novel insight into the dynamics of how CNNs learn brain–relevant visual feature across spatial scale

Chapter 5

Reflections on a NeuroAI Approach to Understanding Infant Visual Development

Infancy is a period of remarkably efficient learning, yet our understanding of brain function during this time remains limited. In this thesis, I propose that NeuroAI has unique potential to provide insights into the development of brain function. I explore this potential through the lens of visual development, a domain with a rich history of revealing general principles of cortical function. To date, the use of NeuroAI approaches for understanding the development of brain function has been limited by the lack of suitable neuroimaging data. Through the Foundations of Cognition project, we have demonstrated that, despite its challenges, awake infant functional MRI is viable at scale, enabling fine-grained characterisation of visual representations along the developing ventral visual stream. Using this rich dataset, I applied advanced computational models – deep neural networks – to investigate the emergence of multiscale object processing in the developing ventral visual stream. I focus on this capacity as its developmental trajectory is of keen interest to cognitive and behavioural scientists studying development and computer scientists seeking to build more robust, shape-based AI visual recognition systems, making it a natural point of convergence between the disciplines. In this chapter, I reflect on what we have learned about developing visual cortex and, in turn, what we have learned about using deep neural networks as models of developing object processing. I discuss limitations of the current work and identify promising avenues for future research.

5.1 General discussion

5.1.1 Insights into developing visual cortex

5.1.1.A Features encoded by developing visual cortex

Deep neural networks were trained on models with various visual diets to gain insights into the features encoded by the developing visual cortex. Using this approach, I demonstrated that the features encoded by infant V1 are richer than previously appreciated, while extrastriate regions exhibit protracted development, with a sustained bias towards coarser object features.

In Chapter 3, I developed a novel framework for characterising the spatial scales of the object features encoded in the ventral visual stream by systematically training and testing DNNs on images with graded levels of blur and comparing the resulting model representations with object responses recorded from the brain. This yielded two-dimensional tuning maps that summarised the spatial scale of the features encoded by a given brain region, enabling comparison across development. This approach was particularly effective at characterising object features in V1 and the extrastriate cortex (V2–V4), which in adults were best modelled by DNNs trained on blurry images and tested on sharp images.

Taking similarity with adults (scaled covariance) as a metric of maturity, infants as young as 2 months were found to have mature patterns of responses in V1, while extrastriate cortex exhibited protracted development, remaining immature through 9 months. In Chapter 4, by isolating low, mid and high spatial frequency features using bandpass

filtering, I subsequently demonstrated that this immaturity in extrastriate regions was accompanied by a relative bias towards encoding coarser features as compared to adults. This indicated a potential inductive bias in the infant brain that persists beyond the development of visual acuity. This was particularly intriguing as coarser information has been shown to aid categorisation (French *et al.*, 2002; Jinsi, Henderson and Tarr, 2023), which could be particularly helpful in supporting basic concept learning over the first year.

In V1, multiscale object feature tuning was found to be qualitatively adult-like by 2 months. Fascinatingly, these visual representations were found to contain object-specific features that corresponded less to a model trained on retinal waves than to a model trained on sharp objects, demonstrating that early visual responses are richer than previously appreciated.

While training solely on retinal waves was a poor fit for V1 across development, retinal-wave pretraining was found to increase alignment with V1 relative to a sharp-trained model, specifically when evaluated on blurred images. This suggested that retinal waves provide valuable pretraining by biasing towards extracting coarser features.

In conjunction, these findings illustrate the potential of DNNs to shed new light on the features encoded by the developing ventral visual stream. In the next sections, I propose possible directions for future work.

5.1.1.A.a Investigating the effect of retinal wave pretraining

A promising direction for future work would be to unpack the benefit conferred by retinal wave pretraining when testing on blurry stimuli. One approach would be to characterise the pretrained model's responses directly, for example by quantifying the network's shape bias (Geirhos *et al.*, 2018a), and probing its internal representations, for example by measuring the size of its first layer receptive fields (Vogelsang *et al.*, 2018). A complementary approach would be to test mechanistic hypotheses as to why retinal wave pretraining increases correspondence with V1. In animals, retinal waves contribute to the activity-dependent establishment and refinement of early visual maps, including retinotopy (Huberman, Feller and Chapman, 2008). It is therefore possible that retinal wave pretraining performs an analogous function in models.

Relatedly, several topographic DNNs have been developed that implement cortex-like wiring constraints (Blauch, Behrmann and Plaut, 2022; Doshi and Konkle, 2023; Margalit *et al.*, 2024; Lu *et al.*, 2025a), and some have demonstrated superior alignment with the high-level category-selective cortex relative to spatially unconstrained models using RSA (e.g. Margalit *et al.*, 2024). An interesting possibility would be to compare the retinal-wave-pretrained model to topographic DNNs to explore whether their spatial smoothness yields similar representational geometry (using model-to-model RSA, as in Chapter 4) and increased correspondence with V1 when tested on blurry images.

5.1.1.A.b Feature visualisation

A particularly appealing future direction will be to use the recent, innovative visualisation methods to gain tangible insights into what the features encoded by the

ventral visual stream might look like, and how this changes across infancy and into adulthood (e.g., Görgün, Schiele and Fischer, 2025).

5.1.1.A.c Method of model–brain comparison

To date, we have investigated model–brain comparisons using classic representational similarity analysis (RSA; Kriegeskorte, Mur and Bandettini, 2008). This approach is typically unweighted – treating all model units as contributing equally when computing representational dissimilarities. However, some units may be more informative than others when modelling the brain, or the units may be present in non-optimal proportions, such that an unweighted comparison underestimates the model fit (Kaniuth and Hebart, 2022). Various methods of reweighting have therefore been proposed and are often used to pull out the best-fitting components of a model's response. The target of this reweighting can be the multivariate pattern of response across voxels in a brain region (Kaniuth and Hebart, 2022; Storrs et al., 2021; Jozwik et al., 2017), or univariate activity at the level of single voxels, known as voxel-wise encoding models (Konkle and Alvarez, 2022; Khaligh-Razavi et al., 2017; Kay et al., 2008). These approaches are typically performed at the individual level. Adapting them to the group level, however, could help reveal whether infant responses correspond to the same components of the model response as adults. Indeed, reweighting at the group level may be less prone to overfitting as the alignment is estimated from more stable, averaged response patterns. This approach could be a particularly useful complement to feature visualisation, as discussed above, by revealing which model components best align with responses at each age.

A recent analysis of more than 200 diverse DNNs varying by architecture and training data, among other properties, revealed that, while there was variation in the representational spaces among the most brain–aligned models, this variation substantially decreased following reweighting, indicating that the brain–aligned components were similar across models (Conwell et al., 2024). An interesting question for future research is whether this is also true across development.

5.1.1.A.d Developing object responses throughout childhood

The protracted development of extrastriate regions, which showed limited change between 2 and 9 months of age, suggests that acquiring data from older children might provide complementary insights. Indeed, there is evidence of prolonged development of object processing into childhood, even in response to familiar classes of stimuli, such as animals (Dekker *et al.*, 2011).

5.1.1.B Investigating learning trajectories and leveraging the full capabilities of NeuroAI

A novel aspect of the current work was that responses from the developing VVS were compared to visual representations from DNNs as they learned to extract object features throughout training. While prior work has compared developing brain responses to the extremes of the learning process (untrained and fully trained; e.g., O'Doherty *et al.*, *in press*), to my knowledge this is the first reported comparison of developing brain responses with a model at multiple stages throughout training. Intriguingly, the shapes of these trajectories — as indexed by layer-wise correspondence throughout the critical training window — were qualitatively similar across development, although the

magnitude of the correspondence differed between groups. Note, however, that a quantitative comparison of these patterns across ages was not performed. The apparent similarity in the patterns of correspondence throughout training could suggest that the models are capturing components of the visual representations that are shared across ages, however, further investigation is required.

Why might the models show similar trajectories of convergence with the early visual cortex (EVC) across development? Intuitively, one might expect the models to converge more quickly with infants, with less-trained models better capturing human representations early in development. One possible explanation for why this pattern was not observed is that the models were trained using a supervised learning algorithm, which is not a plausible model of learning in young infants. Despite extensive visual experience, 9-month-old infants have a very small vocabulary, comprehending an average of 37 words and gestures (Frank *et al.*, 2021). In related work, we have demonstrated that, at the endpoint of training, a CNN trained using a self-supervised learning objective (Konkle and Alvarez, 2022) yielded statistically significant differences, for some model layers, in the magnitude of model-brain correspondence between infants and adults in ventral visual cortex (O'Doherty *et al.*, *in press*). However, differences were not observed for the EVC. An alternative explanation for the converging trajectories could be that the model components that align with brain responses are largely shared across infants and adults. This would cause learning trajectories to appear similar across development. This could be investigated using weighted RSA, as discussed above. A valuable future direction would be to compare developing brain representations with models trained using more developmentally plausible objectives

throughout training, as important developmental differences in model–brain correspondence may emerge.

Indeed, the success of the model–brain comparisons in this work for generating novel insights into the development of the visual cortex motivates further research that leverages the full capabilities of NeuroAI, including variation in learning objective, architecture, and training diet.

5.1.2 Insights for modelling the ventral visual stream with deep neural networks

5.1.2.A Architectural biases

This work also yielded insights into the learning dynamics of DNNs. In Chapter 4, I found that DNNs trained on high spatial frequency (HSF) object information were similar to early visual cortex (EVC) from the outset of training and converged more quickly over the critical training window than models trained on low spatial frequencies. This suggests a potential inductive bias that drives models to preferentially learn HSF features, consistent with prior findings (Doerig *et al.*, 2020). The immediate availability of HSF information could also contribute to the continual learning problem observed when DNNs are trained on progressively sharper images mimicking developing acuity (Jang and Tong, 2021). As discussed, however, converging evidence suggests that while this HSF bias may limit alignment with human behavioural judgments, which tend to prioritise global shape over local texture (Geirhos *et al.*, 2018a), it may not necessarily limit their utility as models of ventral-stream representations (Vogelsang *et al.*, 2018; Ayzenberg and Behrmann, 2022; Loke *et al.*, 2025).

5.1.2.B The potential of critical periods

The distinct developmental trajectories of object tuning across regions of the developing ventral visual stream (VVS; Chapter 3) suggest that the standard practice of modelling VVS development as a parallel optimisation process is an oversimplification. This highlights an important difference between brains and many current models. While DNNs typically include hyperparameters, such as the learning rate, which modulate how quickly they have the capacity to learn, these parameters are typically applied uniformly across the network rather than varying across layers in a developmentally staged way (Vogelsang, Vogelsang and Sinha, 2026). Interestingly, a recent implementation of sequential critical periods has been shown to promote more stable representation learning in DNNs (Delrocq *et al.*, 2024).

Classically, critical periods are characterised as windows of heightened plasticity that occur at particular points in development, with genetic factors playing an important role in their regulation (Hensch, 2004). Experience has also been shown to play an important role, for example, transient monocular deprivation in infancy can lead to poorer long-term outcomes than comparable binocular deprivation (Maurer, 2017). Indeed, critical-period-like dynamics have been shown to arise in neural network models as a consequence of the learning process itself (Elman, 1993; Achille, Rovere and Soatto, 2018), as a model progressively commits to a particular solution, narrowing future possibilities (Elman, 1996).

Considering critical periods might be helpful for future modelling, not only for achieving more biologically inspired vision but also for mitigating continual learning problems, as has been observed with coarse-to-fine training regimes (Jang and Tong, 2021).

5.1.2.C Modelling the trajectory of perceptual development

The methods in the current work involved training models on a single transformation and using the resulting models to probe underlying visual responses. This aided interpretability of the resulting representations, maximised the likelihood of observing an effect, and avoided potential continual learning problems. Given the success of this approach, a natural extension would be to compare developing visual responses against models trained with trajectories of perceptual curricula (e.g., Lu *et al.*, 2025b), which might yield particularly interesting results across training. The current work also highlights several fruitful directions for future modelling of perceptual development, such as incorporating retinal waves.

5.1.3 Clinical potential

Computational modelling has potential to provide insight into systematic differences in information processing that have been observed in some clinical and neurodevelopmental populations (Elman, 2005; Mareschal and Thomas, 2007). For example, individuals diagnosed with Williams syndrome or autism have been reported to rely more on featural than configural information when completing certain tasks (Elman, 2005; Karmiloff-Smith, 1998; Mareschal and Thomas, 2007; Behrmann *et al.*, 2006). A similar behavioural pattern has been reported in individuals born with dense

bilateral congenital cataracts who received sight-restoring surgery after approximately two months of age (Vogelsang *et al.*, 2018; Maurer, Mondloch and Lewis, 2007). Although these groups may show some convergent behavioural signatures, the underlying mechanisms may be distinct (Behrmann *et al.*, 2006). As demonstrated by work on sight restoration, DNNs provide a framework for testing mechanistic hypotheses about what drives such differences in information processing (Vogelsang *et al.*, 2018; Vogelsang *et al.*, 2024b).

Beyond computational modelling, awake infant fMRI has considerable potential for characterising developing brain function in clinical contexts. At present, for example, although many infants who experience perinatal brain injury later develop cognitive or motor difficulties, neonatologists do not currently have a tool to identify which infants are most likely to do so, precluding early intervention. Measuring patterns of functional brain activity may provide unique prognostic information and serve as a valuable tool for clinicians (Linke *et al.*, 2018). We expand on this potential in related work (King et al., in preparation)**.**

5.1.4 Limitations

An inherent limitation in the NeuroAI approach that I adopted is that the biological plausibility of the models is limited. Biological plausibility is not necessary for many of the goals of NeuroAI – for example, it is not relevant if networks are being used as a way to summarise key statistical properties of the environment (e.g., the visual features that best discriminate object classes). However, for other goals, such as using DNNs as models of infant learning, it does limit the degree of possible correspondence. Further,

as noted above, there are several unmodelled factors, such as critical periods, that may be important for learning. Although a strength of these models is their ability to fit training data, there are also inductive biases inherent in the choice of architecture and learning rule, which may or may not reflect biases in the brain. As a whole, these models can provide useful insights, but it is important to remain cognisant of their limitations.

A further limitation of this work is that outside of the laboratory infants generally do not experience objects in isolation but rather embedded in scenes. How infants learn to segment objects from scenes is therefore an important question beyond the scope of the current work. Motion has been hypothesised to play an important role in this process (Braddick and Atkinson, 2011), emphasising the importance of dynamic visual experience – an important point of divergence between the infant learning environment and the static images on which the current class of models is typically trained. Indeed, cortical areas involved in motion processing, such as the middle temporal complex, have been shown to exhibit selective activation in infants as young as 5 weeks (the youngest age tested; Biagi *et al.*, 2015; Biagi *et al.*, 2023). Relatedly, research on neural event segmentation has shown that infants parse their visual experience across time in distinct ways from adults (Yates *et al.*, 2022). How infants parse their visual world in space remains largely to be determined.

While we have demonstrated that awake infant functional MRI is viable at scale, MRI as a neuroimaging modality remains resource intensive – in terms of both personnel time and scanner access costs — limiting scalability to well-resourced laboratories with access to MRI facilities (Margolis *et al.*, 2025).

5.2 A new window into the infant's visual world

This work made several important advances. The FOUNDCOG project constitutes a landmark achievement, as the largest cross-sectional and the first longitudinal study to acquire awake infant fMRI at scale, enabling fine-grained characterisation of developing visual representations in the infant brain. This unique dataset provides a methodological bridge towards leveraging the capabilities of contemporary computational models to understand the developing mind. In this thesis, through an investigation of multiscale object processing, I have demonstrated that such NeuroAI approaches have substantial capacity to provide insights into the developing visual cortex.

Understanding visual development has far-reaching potential for illuminating both the infant and adult mind (Tansey *et al.*, 2024; Ellis and Turk-Browne, 2018), and for shedding new light on long-standing debates in cognitive science, such as the origins of category selectivity in the ventral visual stream (Arcaro and Livingstone, 2021; Blauch, Behrmann and Plaut, 2022; Blauch, Behrmann and Plaut, 2025). Reciprocally, this understanding can inform the development of more robust, human-like computer vision systems. As AI systems become increasingly embedded in daily life, and their associated computational demands continue to grow, infant learning is increasingly appreciated as a source of insights that may serve as a roadmap towards building more efficient and intelligent machines (Zaadnoordijk, Besold and Cusack, 2022).

There is tangible momentum at the intersection of these fields. As cognitive computational neuroscience matures (Naselaris *et al.*, 2018), developmental cognitive computational neuroscience is taking its first steps, as evidenced by a dedicated satellite event at the annual Cognitive Computational Neuroscience conference in 2025 (‘*From Child to Machine Learning: The potential and challenges of translating Developmental Principles into Neural Network Design’*), reflecting increasing interest in these approaches. Emerging advances in fetal, infant and toddler neuroimaging will continue to complement this work, enabling the characterisation of cognitive function in the youngest minds (Pollatou *et al.*, 2022; Spann *et al.*, 2023; Chin *et al.*, 2026). It is my prediction that infant NeuroAI will see substantial growth in the years to come, as it continues to offer a new window into the infant’s visual world.

Bibliography

Achille, A., Rovere, M. and Soatto, S. (2018) ''Critical learning periods in deep networks'', *International Conference on Learning Representations*.

Adams, R. J. and Courage, M. L. (2002) 'Using a single test to measure human contrast sensitivity from early childhood to maturity', *Vision Res,* 42(9), pp. 1205-10.

Adams, R. J., Courage, M. L. and Mercer, M. E. (1991) 'Deficiencies in human neonates' color vision: photoreceptoral and neural explanations', *Behav Brain Res,* 43(2), pp. 109-14.

Agyeman, K., McCarty, T., Multani, H., Mattingly, K., Koziar, K., Chu, J., Liu, C., Kokkoni, E. and Christopoulos, V. (2023) 'Task-based functional neuroimaging in infants: a systematic review', *Front Neurosci,* 17, pp. 1233990.

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., Nau, M., Caron, B., Pestilli, F., Charest, I., Hutchinson, J. B., Naselaris, T. and Kay, K. (2022) 'A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence', *Nat Neurosci,* 25(1), pp. 116-126.

Amunts, K., Mohlberg, H., Bludau, S., Caspers, S., Brandstetter, A., Eickhoff, S. B., Pieperhoff, P., Dickscheid, T. and Zilles, K. (2020a) 'Whole-brain parcellation of the Julich-Brain Cytoarchitectonic Atlas' (Version 2.2). Available at: https://doi.org/10.25493/TAKY-64D (Accessed: 2025).

Amunts, K., Mohlberg, H., Bludau, S. and Zilles, K. (2020b) 'Julich-Brain: A 3D probabilistic atlas of the human brain's cytoarchitecture', *Science,* 369(6506), pp. 988-992.

Anderson, A. W., Marois, R., Colson, E. R., Peterson, B. S., Duncan, C. C., Ehrenkranz, R. A., Schneider, K. C., Gore, J. C. and Ment, L. R. (2001) 'Neonatal auditory activation detected by functional magnetic resonance imaging', *Magn Reson Imaging,* 19(1), pp. 1-5.

Anderson, E. M., Candy, T. R., Gold, J. M. and Smith, L. B. (2024) 'An edge-simplicity bias in the visual input to young infants', *Sci Adv,* 10(19), pp. eadj8571.

Andersson, J. L., Skare, S. and Ashburner, J. (2003) 'How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging', *Neuroimage,* 20(2), pp. 870-88.

Arcaro, M. J. and Livingstone, M. S. (2017) 'A hierarchical, retinotopic proto-organization of the primate visual system at birth', *Elife,* 6.

Arcaro, M. J. and Livingstone, M. S. (2021) 'On the relationship between maps and domains in inferotemporal cortex', *Nat Rev Neurosci,* 22(9), pp. 573-583.

Arcaro, M. J., Schade, P. F., Vincent, J. L., Ponce, C. R. and Livingstone, M. S. (2017) 'Seeing faces is necessary for face-domain formation', *Nat Neurosci,* 20(10), pp. 1404-1412.

Aslin, R. N. (2007) 'What's in a look?', *Dev Sci,* 10(1), pp. 48-53.

Atkinson, J., Braddick, O. and Braddick, F. (1974) 'Acuity and contrast sensivity of infant vision', *Nature,* 247(5440), pp. 403-4.

Avberšek, L. K., Zeman, A. and Op de Beeck, H. (2021) 'Training for object recognition with increasing spatial frequency: A comparison of deep learning with human vision', *J Vis,* 21(10), pp. 14.

Ayzenberg, V. and Behrmann, M. (2022) 'Does the brain's ventral visual pathway compute object shape?', *Trends Cogn Sci,* 26(12), pp. 1119-1132.

Ayzenberg, V. and Behrmann, M. (2024) 'Development of visual object recognition', *Nature Reviews Psychology,* 3(2), pp. 73-90.

Ayzenberg, V., Sener, S. B., Novick, K. and Lourenco, S. F. (2025) 'Fast and robust visual object recognition in young children', *Sci Adv,* 11(27), pp. eads6821.

Azhari, A., Truzzi, A., Neoh, M. J., Balagtas, J. P. M., Tan, H. H., Goh, P. P., Ang, X. A., Setoh, P., Rigo, P., Bornstein, M. H. and Esposito, G. (2020) 'A decade of infant neuroimaging research: What have we learned and where are we going?', *Infant Behav Dev,* 58, pp. 101389.

Baker, N. and Elder, J. H. (2022) 'Deep learning models fail to capture the configural nature of human shape perception', *iScience,* 25(9), pp. 104913.

Baker, N., Lu, H., Erlikhman, G. and Kellman, P. J. (2018) 'Deep convolutional networks do not classify based on global object shape', *PLoS Comput Biol,* 14(12), pp. e1006613.

Banks, M. S. and Bennett, P. J. (1988) 'Optical and photoreceptor immaturities limit the spatial and chromatic vision of human neonates', *J Opt Soc Am A,* 5(12), pp. 2059-79.

Banks, M. S. and Salapatek, P. (1976) 'Contrast sensitivity function of the infant visual system', *Vision Res,* 16(8), pp. 867-9.

Banks, M. S., Stephens, B. R. and Hartmann, E. E. (1985) 'The development of basic mechanisms of pattern vision: spatial frequency channels', *J Exp Child Psychol,* 40(3), pp. 501-27.

Bao, P., She, L., McGill, M. and Tsao, D. Y. (2020) 'A map of object space in primate inferotemporal cortex', *Nature*, pp. 1-6.

Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmid, A. M., Dale, A. M., Hämäläinen, M. S., Marinkovic, K., Schacter, D. L., Rosen, B. R. and Halgren, E. (2006) 'Top-down facilitation of visual recognition', *Proc Natl Acad Sci U S A,* 103(2), pp. 449-54.

Bau, D., Zhu, J. Y., Strobelt, H., Lapedriza, A., Zhou, B. and Torralba, A. (2020) 'Understanding the role of individual units in a deep neural network', *Proc Natl Acad Sci U S A,* 117(48), pp. 30071-30078.

Bayet, L., Zinszer, B. D., Reilly, E., Cataldo, J. K., Pruitt, Z., Cichy, R. M., Nelson, C. A. and Aslin, R. N. (2020) 'Temporal dynamics of visual representations in the infant brain', *Dev Cogn Neurosci,* 45, pp. 100860.

Behm, L., Yates, T. S., Trach, J. E., Choi, D., Du, H., Osumah, C., Deen, B., Kosakowski, H. L., Chen, E. M., Kamps, F. S., Olson, H. A., Ellis, C. T., Saxe, R. and Turk-Browne, N. B. (2026) 'Functional Magnetic Resonance Imaging in Awake Infants: Insights From More Than 750 Scanning Sessions', *Infancy,* 31(1), pp. e70062.

Behrmann, M., Avidan, G., Leonard, G. L., Kimchi, R., Luna, B., Humphreys, K. and Minshew, N. (2006) 'Configural processing in autism and its relationship to face processing', *Neuropsychologia,* 44(1), pp. 110-29.

Ben-Yakov, A. and Henson, R. N. (2018) 'The Hippocampal Film Editor: Sensitivity and Specificity to Event Boundaries in Continuous Experience', *J Neurosci,* 38(47), pp. 10057-10068.

Bergmann, C., Kominsky, J. F., Birules, J., DeBolt, M. and Ward, E. K. (2020) 'Attention Getter'. Available at: https://osf.io/wfe9r (Accessed.

Bethlehem, R. A. I., Seidlitz, J., White, S. R., Vogel, J. W., Anderson, K. M., Adamson, C., Adler, S., Alexopoulos, G. S., Anagnostou, E., Areces-Gonzalez, A., Astle, D. E., Auyeung, B., Ayub, M., Bae, J., Ball, G., Baron-Cohen, S., Beare, R., Bedford, S. A., Benegal, V., ... and VETSA (2022) 'Brain charts for the human lifespan', *Nature,* 604(7906), pp. 525-533.

Biagi, L., Crespi, S. A., Tosetti, M. and Morrone, M. C. (2015) 'BOLD Response Selective to Flow-Motion in Very Young Infants', *PLoS Biol,* 13(9), pp. e1002260.

Biagi, L., Tosetti, M., Crespi, S. A. and Morrone, M. C. (2023) 'Development of BOLD Response to Motion in Human Infants', *J Neurosci,* 43(21), pp. 3825-3837.

Blauch, N. M., Behrmann, M. and Plaut, D. (2025) 'Retinotopic scaffolding of high-level vision', *OSF*.

Blauch, N. M., Behrmann, M. and Plaut, D. C. (2022) 'A connectivity-constrained computational account of topographic organization in primate high-level visual cortex', *Proc Natl Acad Sci U S A,* 119(3).

Blumberg, M. S. and Adolph, K. E. (2023) 'Protracted development of motor cortex constrains rich interpretations of infant cognition', *Trends Cogn Sci,* 27(3), pp. 233-245.

Boden, M. A. (1978) 'Artificial intelligence and Piagetian theory', *Synthese,* 38(3), pp. 389-414.

Bowers, J. S., Malhotra, G., Dujmović, M., Llera Montero, M., Tsvetkov, C., Biscione, V., Puebla, G., Adolfi, F., Hummel, J. E., Heaton, R. F., Evans, B. D., Mitchell, J. and Blything, R. (2022) 'Deep problems with neural network models of human vision', *Behav Brain Sci,* 46, pp. e385.

Bracci, S. and Op de Beeck, H. (2016) 'Dissociations and Associations between Shape and Category Representations in the Two Visual Pathways', *J Neurosci,* 36(2), pp. 432-44.

Bracci, S. and Op de Beeck, H. P. (2023) 'Understanding Human Object Vision: A Picture Is Worth a Thousand Representations', *Annu Rev Psychol,* 74, pp. 113-135.

Braddick, O. and Atkinson, J. (2011) 'Development of human visual function', *Vision Res,* 51(13), pp. 1588-609.

Brown, A. M. (1990) 'Development of visual sensitivity to light and color vision in human infants: a critical review', *Vision Res,* 30(8), pp. 1159-88.

Brown, A. M. and Yamamoto, M. (1986) 'Visual acuity in newborn and preterm infants measured with grating acuity cards', *Am J Ophthalmol,* 102(2), pp. 245-53.

Burkhalter, A. (1993) 'Development of forward and feedback connections between areas V1 and V2 of human visual cortex', *Cereb Cortex,* 3(5), pp. 476-87.

Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M. and Ecker, A. S. (2019) 'Deep convolutional models improve predictions of macaque V1 responses to natural images', *PLoS Comput Biol,* 15(4), pp. e1006897.

Campbell, F. W. and Robson, J. G. (1968) 'Application of Fourier analysis to the visibility of gratings', *J Physiol,* 197(3), pp. 551-66.

Cappell, B. (2023) 'rwave-4096 – Retinal Wave Dataset' (Version 1.0.0) (Accessed.

Cappell, B. and Stoll, A. (2023) 'ANNs pre-trained on Retinal Waves' (Version 3.0.0) (Accessed.

Cappell, B., Stoll, A., Umah, C. W. and Egger, B. 'Reward: Retinal waves for pre-training artificial neural networks mimicking real prenatal development'. PMLR, 76-86.

Chao, L. L., Haxby, J. V. and Martin, A. (1999) 'Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects', *Nat Neurosci,* 2(10), pp. 913-9.

Chapman, B., Gödecke, I. and Bonhoeffer, T. (1999) 'Development of orientation preference in the mammalian visual cortex', *J Neurobiol,* 41(1), pp. 18-24.

Chin, J. H., Wyburd, M. K., Ayzenberg, V., Bayet, L., Bilgic, B., Chen, E. M., Chen, Y., Dineen, Á., Fujita, S., Liu, J., Jun, Y., Camacho, M. C. and Zöllei, L. (2026) 'Deep learning in fetal, infant, and toddler neuroimaging research', *Dev Cogn Neurosci,* 78, pp. 101680.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A. and Oliva, A. (2016) 'Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence', *Sci Rep,* 6, pp. 27755.

Condé, F., Lund, J. S. and Lewis, D. A. (1996) 'The hierarchical development of monkey visual cortical regions as revealed by the maturation of parvalbumin-immunoreactive neurons', *Brain Res Dev Brain Res,* 96(1-2), pp. 261-76.

Author (2022) *Nilearn* (Version 0.9.2).

Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A. and Konkle, T. (2024) 'A large-scale examination of inductive biases shaping high-level visual representation in brains and machines', *Nat Commun,* 15(1), pp. 9383.

Courage, M. L. and Adams, R. J. (1990) 'Visual acuity assessment from birth to three years using the acuity card procedure: cross-sectional and longitudinal samples', *Optom Vis Sci,* 67(9), pp. 713-8.

Courage, M. L. and Adams, R. J. (1996) 'Infant peripheral vision: the development of monocular visual acuity in the first 3 months of postnatal life', *Vision Res,* 36(8), pp. 1207-15.

Cusack, R., McCuaig, O. and Linke, A. C. (2018) 'Methodological challenges in the comparison of infant fMRI across age groups', *Dev Cogn Neurosci,* 33, pp. 194-205.

Cusack, R., O'Doherty, C., Charvet, C. J. and Ranzato, M. (2025) 'Defending the foundation model view of infant development', *Trends Cogn Sci,* 29(7), pp. 589-590.

Cusack, R., Ranzato, M. and Charvet, C. J. (2024) 'Helpless infants are learning a foundation model', *Trends Cogn Sci,* 28(8), pp. 726-738.

Cusack, R., Wild, C., Linke, A. C., Arichi, T., Lee, D. S. and Han, V. K. (2015) 'Optimizing stimulation and analysis protocols for neonatal fMRI', *PLoS One,* 10(8), pp. e0120202.

Cusack, R., Wild, C. J., Zubiaurre-Elorza, L. and Linke, A. C. (2018) 'Why does language not emerge until the second year?', *Hear Res,* 366, pp. 75-81.

Dall'Orso, S., Steinweg, J., Allievi, A. G., Edwards, A. D., Burdet, E. and Arichi, T. (2018) 'Somatotopic Mapping of the Developing Sensorimotor Cortex in the Preterm Human Brain', *Cereb Cortex,* 28(7), pp. 2507-2515.

De Valois, R. L. and De Valois, K. K. (1980) 'Spatial vision', *Annu Rev Psychol,* 31, pp. 309-41.

Dean, D. C., Tisdall, M. D., Wisnowski, J. L., Feczko, E., Gagoski, B., Alexander, A. L., Edden, R. A. E., Gao, W., Hendrickson, T. J., Howell, B. R., Huang, H., Humphreys, K. L., Riggins, T., Sylvester, C. M., Weldon, K. B., Yacoub, E., Ahtam, B., Beck, N., Banerjee, S., ... and Group, H. M. W. (2024) 'Quantifying brain development in the HEALthy Brain and Child Development (HBCD) Study: The magnetic resonance imaging and spectroscopy protocol', *Dev Cogn Neurosci,* 70, pp. 101452.

Deen, B., Richardson, H., Dilks, D. D., Takahashi, A., Keil, B., Wald, L. L., Kanwisher, N. and Saxe, R. (2017) 'Organization of high-level visual cortex in human infants', *Nat Commun,* 8, pp. 13995.

Dehaene-Lambertz, G., Dehaene, S. and Hertz-Pannier, L. (2002) 'Functional neuroimaging of speech perception in infants', *Science,* 298(5600), pp. 2013-5.

Dekker, T., Mareschal, D., Sereno, M. I. and Johnson, M. H. (2011) 'Dorsal and ventral stream activation and object recognition performance in school-age children', *Neuroimage,* 57(3), pp. 659-70.

Dekker, T. M., Farahbakhsh, M., Atkinson, J., Braddick, O. J. and Jones, P. R. (2020) 'Development of the spatial contrast sensitivity function (CSF) during childhood: Analysis of previous findings and new psychophysical data', *J Vis,* 20(13), pp. 4.

Delrocq, A., Wu, Z., Bellec, G. and Gerstner, W. (2024) 'Critical periods support representation learning in a model of cortical processing', *bioRxiv*, pp. 2024.12.20.629674.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L. 'Imagenet: A large-scale hierarchical image database'. 2009: Ieee, 248-255.

DiCarlo, J. J. and Cox, D. D. (2007) 'Untangling invariant object recognition', *Trends Cogn Sci,* 11(8), pp. 333-41.

Dobson, V. and Teller, D. Y. (1978) 'Visual acuity in human infants: a review and comparison of behavioral and electrophysiological studies', *Vision Res,* 18(11), pp. 1469-83.

Doerig, A., Bornet, A., Choung, O. H. and Herzog, M. H. (2020) 'Crowding reveals fundamental differences in local vs. global processing in humans and machines', *Vision Res,* 167, pp. 39-45.

Doerig, A., Sommers, R. P., Seeliger, K., Richards, B., Ismael, J., Lindsay, G. W., Kording, K. P., Konkle, T., van Gerven, M. A. J., Kriegeskorte, N. and Kietzmann, T. C. (2023) 'The neuroconnectionist research programme', *Nat Rev Neurosci,* 24(7), pp. 431-450.

Dominguez, M. and Jacobs, R. A. (2003) 'Developmental constraints aid the acquisition of binocular disparity sensitivities', *Neural Comput,* 15(1), pp. 161-82.

Doshi, F. R., Fel, T., Konkle, T. and Alvarez, G. (2025) 'Visual Anagrams Reveal Hidden Differences in Holistic Shape Processing Across Vision Models', *arXiv preprint arXiv:2507.00493*.

Doshi, F. R. and Konkle, T. (2023) 'Cortical topographic motifs emerge in a self-organized map of object space', *Sci Adv,* 9(25), pp. eade8187.

Downing, P. E., Chan, A. W., Peelen, M. V., Dodds, C. M. and Kanwisher, N. (2006) 'Domain specificity in visual cortex', *Cereb Cortex,* 16(10), pp. 1453-61.

Downing, P. E., Jiang, Y., Shuman, M. and Kanwisher, N. (2001) 'A cortical area selective for visual processing of the human body', *Science,* 293(5539), pp. 2470-3.

Ellis, C. T., Skalaban, L. J., Yates, T. S., Bejjanki, V. R., Córdova, N. I. and Turk-Browne, N. B. (2020) 'Re-imagining fMRI for awake behaving infants', *Nat Commun,* 11(1), pp. 4523.

Ellis, C. T., Skalaban, L. J., Yates, T. S. and Turk-Browne, N. B. (2021a) 'Attention recruits frontal cortex in human infants', *Proc Natl Acad Sci U S A,* 118(12).

Ellis, C. T. and Turk-Browne, N. B. (2018) 'Infant fMRI: A Model System for Cognitive Neuroscience', *Trends Cogn Sci,* 22(5), pp. 375-387.

Ellis, C. T., Yates, T. S., Arcaro, M. J. and Turk-Browne, N. (2025) 'Movies reveal the fine-grained organization of infant visual cortex', *Elife,* 12.

Ellis, C. T., Yates, T. S., Skalaban, L. J., Bejjanki, V. R., Arcaro, M. J. and Turk-Browne, N. B. (2021b) 'Retinotopic organization of visual cortex in human infants', *Neuron,* 109(16), pp. 2616-2626.e6.

Elman, J. L. (1993) 'Learning and development in neural networks: the importance of starting small', *Cognition,* 48(1), pp. 71-99.

Elman, J. L. (1996) *Rethinking innateness : a connectionist perspective on development. Neural network modeling and connectionism* Cambridge, Mass.: MIT Press.

Elman, J. L. (2005) 'Connectionist models of cognitive development: where next?', *Trends Cogn Sci,* 9(3), pp. 111-7.

Emberson, L. L., Crosswhite, S. L., Richards, J. E. and Aslin, R. N. (2017a) 'The Lateral Occipital Cortex Is Selective for Object Shape, Not Texture/Color, at Six Months', *J Neurosci,* 37(13), pp. 3698-3703.

Emberson, L. L., Zinszer, B. D., Raizada, R. D. S. and Aslin, R. N. (2017b) 'Decoding the infant mind: Multivariate pattern analysis (MVPA) using fNIRS', *PLoS One,* 12(4), pp. e0172500.

Epstein, R. and Kanwisher, N. (1998) 'A cortical representation of the local visual environment', *Nature,* 392(6676), pp. 598-601.

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A. and Gorgolewski, K. J. (2019) 'fMRIPrep: a robust preprocessing pipeline for functional MRI', *Nat Methods,* 16(1), pp. 111-116.

Eyre, M., Fitzgibbon, S. P., Ciarrusta, J., Cordero-Grande, L., Price, A. N., Poppe, T., Schuh, A., Hughes, E., O'Keeffe, C., Brandon, J., Cromb, D., Vecchiato, K., Andersson, J., Duff, E. P., Counsell, S. J., Smith, S. M., Rueckert, D., Hajnal, J. V., Arichi, T., ... and Edwards, A. D. (2021) 'The Developing Human Connectome Project: typical and disrupted perinatal functional connectivity', *Brain,* 144(7), pp. 2199-2213.

Fair, D. A., Dosenbach, N. U. F., Moore, A. H., Satterthwaite, T. D. and Milham, M. P. (2021) 'Developmental cognitive neuroscience in the era of networks and big data: strengths, weaknesses, opportunities, and threats', *Annual Review of Developmental Psychology,* 3, pp. 249-275.

Fantz, R. L. (1961) 'The origin of form perception', *Sci Am,* 204, pp. 66-72.

Fantz, R. L. (1964) 'Visual Experience in Infants: Decreased Attention to Familiar Patterns Relative to Novel Ones', *Science,* 146(3644), pp. 668-70.

Farhang, E., Toosi, R., Karami, B., Koushki, R., Kheirkhah, N., Shakerian, F., Noroozi, J., Rezayat, E., Vahabie, A. H. and Dehaqani, M. A. (2025) 'The impact of spatial frequency on hierarchical category representation in macaque temporal cortex', *Commun Biol,* 8(1), pp. 801.

Farzin, F., Rivera, S. M. and Whitney, D. (2010) 'Spatial resolution of conscious visual perception in infants', *Psychol Sci,* 21(10), pp. 1502-9.

Fausey, C. M., Jayaraman, S. and Smith, L. B. (2016) 'From faces to hands: Changing visual input in the first two years', *Cognition,* 152, pp. 101-107.

Field, D. J. (1987) 'Relations between the statistics of natural images and the response properties of cortical cells', *J Opt Soc Am A,* 4(12), pp. 2379-94.

Fiorentini, A., Pirchio, M. and Spinelli, D. (1983) 'Electrophysiological evidence for spatial frequency selective mechanisms in adults and infants', *Vision Res,* 23(2), pp. 119-27.

Fitzgibbon, S. P., Harrison, S. J., Jenkinson, M., Baxter, L., Robinson, E. C., Bastiani, M., Bozek, J., Karolis, V., Cordero Grande, L., Price, A. N., Hughes, E., Makropoulos, A., Passerat-Palmbach, J., Schuh, A., Gao, J., Farahibozorg, S. R., O'Muircheartaigh, J., Ciarrusta, J., O'Keeffe, C., ... and Andersson, J. (2020) 'The developing Human Connectome Project (dHCP) automated resting-state functional processing framework for newborn infants', *Neuroimage,* 223, pp. 117303.

Fonov, V., Evans, A. C., Botteron, K., Almli, C. R., McKinstry, R. C., Collins, D. L. and Group, B. D. C. (2011) 'Unbiased average age-appropriate atlases for pediatric studies', *Neuroimage,* 54(1), pp. 313-27.

Frank, M. C. (2023) 'Bridging the data gap between children and large language models', *Trends Cogn Sci,* 27(11), pp. 990-992.

Frank, M. C., Braginsky, M., Yurovsky, D. and Marchman, V. A. (2021) *Variability and consistency in early language learning: The Wordbank project.* MIT Press.

Fransson, P., Skiöld, B., Horsch, S., Nordell, A., Blennow, M., Lagercrantz, H. and Aden, U. (2007) 'Resting-state networks in the infant brain', *Proc Natl Acad Sci U S A,* 104(39), pp. 15531-6.

French, R. M., Mermillod, M., Chauvin, A., Quinn, P. C. and Mareschal, D. 'The importance of starting blurry: Simulating improved basic-level category learning in infants due to weak visual acuity'. 2002.

Gao, W., Lin, W., Grewen, K. and Gilmore, J. H. (2017) 'Functional Connectivity of the Infant Human Brain: Plastic and Modifiable', *Neuroscientist,* 23(2), pp. 169-184.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A. and Brendel, W. (2018a) 'ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness', *arXiv preprint arXiv:1811.12231*.

Geirhos, R., Temme, C. R. M., Rauber, J., Schütt, H. H., Bethge, M. and Wichmann, F. A. 'Generalisation in humans and deep neural networks', 2018, 7538-7550.

Gilmore, J. H., Knickmeyer, R. C. and Gao, W. (2018) 'Imaging structural and functional brain development in early childhood', *Nat Rev Neurosci,* 19(3), pp. 123-137.

Goodale, M. A. and Milner, A. D. (1992) 'Separate visual pathways for perception and action', *Trends Neurosci,* 15(1), pp. 20-5.

Goodfellow, I., Bengio, Y. and Courville, A. (2016) *Deep learning.* MIT press.

Gopnik, A., Meltzoff, A. N. and Kuhl, P. K. (1999) *The scientist in the crib : minds, brains, and how children learn.* 1st edn. New York: Morrow.

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L. and Ghosh, S. S. (2011) 'Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python', *Front Neuroinform,* 5, pp. 13.

Greene, D. J., Koller, J. M., Hampton, J. M., Wesevich, V., Van, A. N., Nguyen, A. L., Hoyt, C. R., McIntyre, L., Earl, E. A., Klein, R. L., Shimony, J. S., Petersen, S. E., Schlaggar, B. L., Fair, D. A. and Dosenbach, N. U. F. (2018) 'Behavioral interventions for reducing head motion during MRI scans in children', *Neuroimage,* 171, pp. 234-245.

Grootswagers, T., Quek, G. L., Zeng, Z. and Varlet, M. (2025) 'Human infant EEG recordings for 200 object images presented in rapid visual streams', *Scientific Data,* 12(1), pp. 407.

Görgün, A., Schiele, B. and Fischer, J. 'VITAL: More Understandable Feature Visualization through Distribution Alignment and Relevant Information Flow'. 2025, 4403-4412.

Güçlü, U. and van Gerven, M. A. (2015) 'Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream', *J Neurosci,* 35(27), pp. 10005-14.

Hajnal, J. V., Myers, R., Oatridge, A., Schwieso, J. E., Young, I. R. and Bydder, G. M. (1994) 'Artifacts due to stimulus correlated motion in functional imaging of the brain', *Magn Reson Med,* 31(3), pp. 283-91.

Hamilton, R., Bach, M., Heinrich, S. P., Hoffmann, M. B., Odom, J. V., McCulloch, D. L. and Thompson, D. A. (2021) 'VEP estimation of visual acuity: a systematic review', *Doc Ophthalmol,* 142(1), pp. 25-74.

Harris, L., Atkinson, J. and Braddick, O. (1976) 'Visual contrast sensitivity of a 6-month-old infant measured by the evoked potential', *Nature,* 264(5586), pp. 570-1.

He, K., Zhang, X., Ren, S. and Sun, J. (2015
) 'Deep residual learning for image recognition'. 2016, 770-778.

Heinke, D., Wachman, P., van Zoest, W. and Leek, E. C. (2021) 'A failure to learn object shape geometry: Implications for convolutional neural networks as plausible models of biological vision', *Vision Res,* 189, pp. 81-92.

Heitmann, C., Zhan, M., Linke, M., Hölig, C., Kekunnaya, R., van Hoof, R., Goebel, R. and Röder, B. (2023) 'Early visual experience refines the retinotopic organization within and across visual cortical regions', *Curr Biol,* 33(22), pp. 4950-4959.e4.

Henderson, M. M., Tarr, M. J. and Wehbe, L. (2023) 'A Texture Statistics Encoding Model Reveals Hierarchical Feature Selectivity across Human Visual Cortex', *J Neurosci,* 43(22), pp. 4144-4161.

Hendrickson, A. and Boothe, R. (1976) 'Morphology of the retina and dorsal lateral geniculate nucleus in dark-reared monkeys (Macaca nemestrina)', *Vision research,* 16(5), pp. 517-IN5.

Hendrix, C. L. and Thomason, M. E. (2022) 'A survey of protocols from 54 infant and toddler neuroimaging research labs', *Dev Cogn Neurosci,* 54, pp. 101060.

Hensch, T. K. (2004) 'Critical period regulation', *Annu Rev Neurosci,* 27, pp. 549-79.

Hinton, G. E., McClelland, J. L. and Rumelhart, D. E. (1986) 'Distributed Representations', in Rumelhart, D.E., McClelland, J.L. and Group, P.R. (eds.) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations*: MIT Press, pp. 77–109.

Hirsch, H. V. and Spinelli, D. N. (1970) 'Visual experience modifies distribution of horizontally and vertically oriented receptive fields in cats', *Science,* 168(3933), pp. 869-71.

Howell, B. R., Styner, M. A., Gao, W., Yap, P. T., Wang, L., Baluyot, K., Yacoub, E., Chen, G., Potts, T., Salzwedel, A., Li, G., Gilmore, J. H., Piven, J., Smith, J. K., Shen, D., Ugurbil, K., Zhu, H., Lin, W. and Elison, J. T. (2019) 'The UNC/UMN Baby Connectome Project (BCP): An overview of the study design and protocol development', *Neuroimage,* 185, pp. 891-905.

Hubel, D. H. and Wiesel, T. N. (1959) 'Receptive fields of single neurones in the cat's striate cortex', *The Journal of physiology,* 148(3), pp. 574-591.

Hubel, D. H. and Wiesel, T. N. (1962) 'Receptive fields, binocular interaction and functional architecture in the cat's visual cortex', *J Physiol,* 160(1), pp. 106-54.

Huberman, A. D., Feller, M. B. and Chapman, B. (2008) 'Mechanisms underlying development of visual maps and receptive fields', *Annu Rev Neurosci,* 31, pp. 479-509.

Huth, A. G., Nishimoto, S., Vu, A. T. and Gallant, J. L. (2012) 'A continuous semantic space describes the representation of thousands of object and action categories across the human brain', *Neuron,* 76(6), pp. 1210-24.

Huttenlocher, P. R., de Courten, C., Garey, L. J. and Van der Loos, H. (1982) 'Synaptogenesis in human visual cortex--evidence for synapse elimination during normal development', *Neurosci Lett,* 33(3), pp. 247-52.

Imam, N. and L Finlay, B. (2020) 'Self-organization of cortical areas in the development and evolution of neocortex', *Proc Natl Acad Sci U S A,* 117(46), pp. 29212-29220.

Jacob, G., Pramod, R. T., Katti, H. and Arun, S. P. (2021) 'Qualitative similarities and differences in visual object representations between brains and deep networks', *Nature Communications,* 12(1), pp. 1872.

Jagadeesh, A. V. and Gardner, J. L. (2022) 'Texture-like representation of objects in human visual cortex', *Proc Natl Acad Sci U S A,* 119(17), pp. e2115302119.

Jang, H. and Tong, F. (2021) 'Convolutional neural networks trained with a developmental sequence of blurry to clear images reveal core differences between face and object processing', *J Vis,* 21(12), pp. 6.

Jang, H. and Tong, F. (2024) 'Improved modeling of human vision by incorporating robustness to blur in convolutional neural networks', *Nat Commun,* 15(1), pp. 1989.

Jayaraman, S., Fausey, C. M. and Smith, L. B. (2015) 'The Faces in Infant-Perspective Scenes Change over the First Year of Life', *PLoS One,* 10(5), pp. e0123780.

Jenkinson, M., Bannister, P., Brady, M. and Smith, S. (2002) 'Improved optimization for the robust and accurate linear registration and motion correction of brain images', *Neuroimage,* 17(2), pp. 825-41.

Jenkinson, M. and Smith, S. (2001) 'A global optimisation method for robust affine registration of brain images', *Med Image Anal,* 5(2), pp. 143-56.

Jessen, S., Fiedler, L., Münte, T. F. and Obleser, J. (2019) 'Quantifying the individual auditory and visual brain response in 7-month-old infants watching a brief cartoon movie', *Neuroimage,* 202, pp. 116060.

Jinsi, O., Henderson, M. M. and Tarr, M. J. (2023) 'Early experience with low-pass filtered images facilitates visual category learning in a neural network model', *PLoS One,* 18(1), pp. e0280145.

Jozwik, K. M., Kriegeskorte, N., Storrs, K. R. and Mur, M. (2017) 'Deep Convolutional Neural Networks Outperform Feature-Based But Not Categorical Models in Explaining Object Similarity Judgments', *Front Psychol,* 8, pp. 1726.

Kaniuth, P. and Hebart, M. N. (2022) 'Feature-reweighted representational similarity analysis: A method for improving the fit between computational models, brains, and behavior', *Neuroimage,* 257, pp. 119294.

Kanwisher, N., Khosla, M. and Dobs, K. (2023) 'Using artificial neural networks to ask 'why' questions of minds and brains', *Trends Neurosci,* 46(3), pp. 240-254.

Kanwisher, N., McDermott, J. and Chun, M. M. (1997) 'The fusiform face area: a module in human extrastriate cortex specialized for face perception', *J Neurosci,* 17(11), pp. 4302-11.

Karmiloff-Smith, A. (1998) 'Development itself is the key to understanding developmental disorders', *Trends Cogn Sci,* 2(10), pp. 389-98.

Kay, K. N., Naselaris, T., Prenger, R. J. and Gallant, J. L. (2008) 'Identifying natural images from human brain activity', *Nature,* 452(7185), pp. 352-5.

Kayser, C., Körding, K. P. and König, P. (2004) 'Processing of complex stimuli and natural scenes in the visual cortex', *Curr Opin Neurobiol,* 14(4), pp. 468-73.

Khaligh-Razavi, S. M., Henriksson, L., Kay, K. and Kriegeskorte, N. (2017) 'Fixed versus mixed RSA: Explaining visual representations by fixed and mixed feature sets from shallow and deep computational models', *J Math Psychol,* 76(Pt B), pp. 184-197.

Khaligh-Razavi, S. M. and Kriegeskorte, N. (2014) 'Deep supervised, but not unsupervised, models may explain IT cortical representation', *PLoS Comput Biol,* 10(11), pp. e1003915.

Kiat, J. E., Luck, S. J., Beckner, A. G., Hayes, T. R., Pomaranski, K. I., Henderson, J. M. and Oakes, L. M. (2022) 'Linking patterns of infant eye movements to a neural network model of the ventral stream using representational similarity analysis', *Dev Sci,* 25(1), pp. e13155.

Kietzmann, T. C., McClure, P. and Kriegeskorte, N. 2019. Deep Neural Networks in Computational Neuroscience. *Oxford Research Encyclopedia of Neuroscience.* Oxford: Oxford University Press.

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K. A., Cichy, R. M., Hauk, O. and Kriegeskorte, N. (2019) 'Recurrence is required to capture the representational dynamics of the human visual system', *Proc Natl Acad Sci U S A,* 116(43), pp. 21854-21863.

King, G., Dineen, Á. T., O'Doherty, C., Truzzi, A., Caldinelli, C., D'Arcy, E.-L., White, J., Kravchenko, A., Zaadnoordijk, L., Burke, A., Healy, M., Joseph, S., Molloy, E. J., Foran, A., Tarrant, A., Byrne, A. T. and Cusack, R. *in preparation*. Practicalities of scanning an Awake Infant functional MRI cohort.

King, G., Truzzi, A. and Cusack, R. (2023) 'The confound of head position in within-session connectome fingerprinting in infants', *Neuroimage,* 265, pp. 119808.

Kiorpes, L. (2015) 'Visual development in primates: Neural mechanisms and critical periods', *Dev Neurobiol,* 75(10), pp. 1080-90.

Kiorpes, L. (2016) 'The Puzzle of Visual Development: Behavior and Neural Limits', *J Neurosci,* 36(45), pp. 11384-11393.

Knoblauch, K., Vital-Durand, F. and Barbur, J. L. (2001) 'Variation of chromatic sensitivity across the life span', *Vision Res,* 41(1), pp. 23-36.

Konkle, T. and Alvarez, G. A. (2022) 'A self-supervised domain-general learning framework for human ventral stream representation', *Nat Commun,* 13(1), pp. 491.

Konkle, T. and Caramazza, A. (2013) 'Tripartite organization of the ventral stream by animacy and object size', *J Neurosci,* 33(25), pp. 10235-42.

Kosakowski, H. L., Cohen, M. A., Takahashi, A., Keil, B., Kanwisher, N. and Saxe, R. (2022) 'Selective responses to faces, scenes, and bodies in the ventral visual pathway of infants', *Curr Biol,* 32(2), pp. 265-274.e5.

Kravitz, D. J., Peng, C. S. and Baker, C. I. (2011) 'Real-world scene representations in high-level visual cortex: it's the spaces more than the places', *J Neurosci,* 31(20), pp. 7322-33.

Kriegeskorte, N. (2015) 'Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing', *Annu Rev Vis Sci,* 1, pp. 417-446.

Kriegeskorte, N., Mur, M. and Bandettini, P. A. (2008) 'Representational similarity analysis-connecting the branches of systems neuroscience', *Frontiers in systems neuroscience,* 2, pp. 249.

Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012) 'Imagenet classification with deep convolutional neural networks', *Advances in neural information processing systems,* 25, pp. 1097-1105.

Kubilius, J., Bracci, S. and Op de Beeck, H. P. (2016) 'Deep Neural Networks as a Computational Model for Human Shape Sensitivity', *PLoS Comput Biol,* 12(4), pp. e1004896.

Lamme, V. (2014) 'The crack of dawn: perceptual functions and neural mechanisms that mark the transition from unconscious processing to conscious vision', *Open mind*: Open MIND. Frankfurt am Main: MIND Group.

Landau, B., Smith, L. B. and Jones, S. S. (1988) 'The importance of shape in early lexical learning', *Cognitive development,* 3(3), pp. 299-321.

Leat, S. J., Yadav, N. K. and Irving, E. L. (2009) 'Development of visual acuity and contrast sensitivity in children', *Journal of optometry,* 2(1), pp. 19-26.

LeCun, Y., Bengio, Y. and Hinton, G. (2015) 'Deep learning', *Nature,* 521(7553), pp. 436-44.

Lee, T. S., Mumford, D., Zhu, S. C. and Lamme, V. A. F. (1997) 'The role of V1 in shape representation', *Computational Neuroscience: Trends in Research, 1997*: Springer, pp. 697-703.

Li, Z., Ortega Caro, J., Rusak, E., Brendel, W., Bethge, M., Anselmi, F., Patel, A. B., Tolias, A. S. and Pitkow, X. (2023) 'Robust deep learning object recognition models rely on low frequency information in natural images', *PLoS Comput Biol,* 19(3), pp. e1010932.

Linke, A. C. and Cusack, R. (2015) 'Flexible information coding in human auditory cortex during perception, imagery, and STM of complex sounds', *J Cogn Neurosci,* 27(7), pp. 1322-33.

Linke, A. C., Wild, C., Zubiaurre-Elorza, L., Herzmann, C., Duffy, H., Han, V. K., Lee, D. S. C. and Cusack, R. (2018) 'Disruption to functional networks in neonates with perinatal brain injury predicts motor skills at 8 months', *NeuroImage: Clinical,* 18, pp. 399-406.

Linsley, D. and Serre, T. (2023) 'Fixing the problems of deep neural networks will require better training data and learning algorithms', *Behav Brain Sci,* 46, pp. e400.

Loke, J., Soerensen, L. K. A., Groen, I. I. A., Cappaert, N. and Scholte, H. S. (2025) 'Shared texture-like representations, not global form, underlie deep neural network alignment with human visual processing', *bioRxiv*, pp. 2025-08.

Long, B., Sparks, R. Z., Xiang, V., Stojanov, S., Yin, Z., Keene, G. E., Tan, A. W. M., Feng, S. Y., Zhuang, C. and Marchman, V. A. (2024) 'The BabyView dataset: High-resolution egocentric videos of infants' and young children's everyday experiences', *arXiv preprint arXiv:2406.10447*.

Long, B., Yu, C.-P. and Konkle, T. (2018) 'Mid-level visual features underlie the high-level categorical organization of the ventral stream', *Proceedings of the National Academy of Sciences,* 115(38), pp. E9015-E9024.

Lu, Z., Doerig, A., Bosch, V., Krahmer, B., Kaiser, D., Cichy, R. M. and Kietzmann, T. C. (2025a) 'End-to-end topographic networks as models of cortical map formation and human visual behaviour', *Nat Hum Behav*.

Lu, Z., Thorat, S., Cichy, R. M. and Kietzmann, T. C. (2025b) 'Adopting a human developmental visual diet yields robust, shape-based AI vision', *arXiv preprint arXiv:2507.03168*.

Marchal, A., Melchior, M., Dufour, A., Poisbeau, P., Zores, C. and Kuhn, P. (2021) 'Pain Behavioural Response to Acoustic and Light Environmental Changes in Very Preterm Infants', *Children (Basel),* 8(12).

Mareschal, D., French, R. M. and Quinn, P. C. (2000) 'A connectionist account of asymmetric category learning in early infancy', *Dev Psychol,* 36(5), pp. 635-645.

Mareschal, D. and Quinn, P. C. (2001) 'Categorization in infancy', *Trends Cogn Sci,* 5(10), pp. 443-450.

Mareschal, D. and Thomas, M. S. C. (2007) 'Computational modeling in developmental psychology', *IEEE Transactions on Evolutionary Computation,* 11(2), pp. 137-150.

Margalit, E., Lee, H., Finzi, D., DiCarlo, J. J., Grill-Spector, K. and Yamins, D. L. K. (2024) 'A unifying framework for functional organization in early and higher ventral visual cortex', *Neuron,* 112(14), pp. 2435-2451 e7.

Margolis, E. T., Nelson, P. M., Fiske, A., Champaud, J. L. Y., Olson, H. A., Gomez, M. J. C., Dineen, Á., Bulgarelli, C., Troller-Renfree, S. V., Donald, K. A., Spann, M. N., Howell, B., Scheinost, D. and Korom, M. (2025) 'Modality-level obstacles and initiatives to improve representation in fetal, infant, and toddler neuroimaging research samples', *Dev Cogn Neurosci,* 72, pp. 101505.

Matkovic, K., Neumann, L., Neumann, A., Psik, T. and Purgathofer, W. 'Global contrast factor-a new approach to image contrast'. 159-167.

Maule, J., Skelton, A. E. and Franklin, A. (2023) 'The Development of Color Perception and Cognition', *Annu Rev Psychol,* 74, pp. 87-111.

Maurer, D. (2017) 'Critical periods re-examined: evidence from children treated for dense cataracts', *Cognitive Development,* 42, pp. 27-36.

Maurer, D., Lewis, T. L., Brent, H. P. and Levin, A. V. (1999) 'Rapid improvement in the acuity of infants after visual input', *Science,* 286(5437), pp. 108-110.

Maurer, D. and Maurer, C. (1988) *The world of the newborn.* New York: Basic Books.

Maurer, D., Mondloch, C. J. and Lewis, T. L. (2007) 'Sleeper effects', *Dev Sci,* 10(1), pp. 40-7.

May, L., Dauphin, A. and Gjorgjieva, J. (2025) 'Pre-training artificial neural networks with spontaneous retinal activity improves motion prediction in natural scenes', *PLoS Comput Biol,* 21(3), pp. e1012830.

Mazziotta, J., Toga, A., Evans, A., Fox, P., Lancaster, J., Zilles, K., Woods, R., Paus, T., Simpson, G., Pike, B., Holmes, C., Collins, L., Thompson, P., MacDonald, D., Iacoboni, M., Schormann, T., Amunts, K., Palomero-Gallagher, N., Geyer, S., ... and Mazoyer, B. (2001) 'A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM)', *Philos Trans R Soc Lond B Biol Sci,* 356(1412), pp. 1293-322.

McCulloch, W. S. and Pitts, W. (1943) 'A logical calculus of the ideas immanent in nervous activity', *Bulletin of Mathematical Biophysics,* 5(4), pp. 115–133.

Mehrer, J., Spoerer, C. J., Kriegeskorte, N. and Kietzmann, T. C. (2020) 'Individual differences among deep neural network models', *Nat Commun,* 11(1), pp. 5725.

Miao, H. Y. and Tong, F. (2024) 'Convolutional neural network models applied to neuronal responses in macaque V1 reveal limited nonlinear processing', *J Vis,* 24(6), pp. 1.

Mishkin, M., Ungerleider, L. G. and Macko, K. A. (1983) 'Object vision and spatial vision: two cortical pathways', *Trends in neurosciences,* 6, pp. 414-417.

Mongerson, C. R. L., Jennings, R. W., Borsook, D., Becerra, L. and Bajic, D. (2017) 'Resting-State Functional Connectivity in the Infant Brain: Methods, Pitfalls, and Potentiality', *Front Pediatr,* 5, pp. 159.

Mundinano, I. C., Kwan, W. C. and Bourne, J. A. (2015) 'Mapping the mosaic sequence of primate visual cortical development', *Front Neuroanat,* 9, pp. 132.

Müller, N., Snoek, C. G. M., Groen, I. I. A. and Scholte, H. S. (2024) 'Object-zoomed training of convolutional neural networks inspired by toddler development improves shape bias', *bioRxiv*, pp. 2024-05.

Nakashima, Y., Kanazawa, S. and Yamaguchi, M. K. (2021) 'Perception of invisible masked objects in early infancy', *Proc Natl Acad Sci U S A,* 118(27).

Naselaris, T., Bassett, D. S., Fletcher, A. K., Kording, K., Kriegeskorte, N., Nienborg, H., Poldrack, R. A., Shohamy, D. and Kay, K. (2018) 'Cognitive Computational Neuroscience: A New Conference for an Emerging Discipline', *Trends Cogn Sci,* 22(5), pp. 365-367.

Nelson, C. A., Frankeberger, J. and Chambers, C. D. (2024) 'An introduction to the HEALthy Brain and Child Development Study (HBCD) study', *Dev Cogn Neurosci,* 69, pp. 101441.

Nguyen, A., Yosinski, J. and Clune, J. 'Deep neural networks are easily fooled: High confidence predictions for unrecognizable images'. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 427-436.

Nielsen, A. N., Graham, A. M. and Sylvester, C. M. (2023) 'Baby Brains at Work: How Task-Based Functional Magnetic Resonance Imaging Can Illuminate the Early Emergence of Psychiatric Risk', *Biol Psychiatry,* 93(10), pp. 880-892.

Norcia, A. M. and Tyler, C. W. (1985) 'Spatial frequency sweep VEP: visual acuity during the first year of life', *Vision Res,* 25(10), pp. 1399-408.

Norcia, A. M., Tyler, C. W. and Hamer, R. D. (1990) 'Development of contrast sensitivity in the human infant', *Vision Res,* 30(10), pp. 1475-86.

O'Doherty, C., Dineen, Á., Truzzi, A., King, G., Zaadnoordijk, L., Harrison, K., D'Arcy, E. L., White, J., Caldinelli, C., Holloway, T., Kravchenko, A., Diedrichsen, J., Tarrant, A., Byrne, A. T., Foran, A., Molloy, E. J. and Cusack, R. (2026) 'Infants have rich visual categories in ventrotemporal cortex at 2 months of age', *Nat Neurosci*.

Olson, H. A., Camacho, M. C., Abdurokhmonova, G., Ahmad, S., Chen, E. M., Chung, H., Lorenzo, R. D., Dineen, Á., Ganz, M., Licandro, R., Magnain, C., Marrus, N., McCormick, S. A., Rutter, T. M., Wagner, L., Woodruff Carr, K., Zöllei, L., Vaughn, K. A. and Madsen, K. S. (2025) 'Measuring and interpreting individual differences in fetal, infant, and toddler neurodevelopment', *Dev Cogn Neurosci,* 73, pp. 101539.

O'Doherty, C., Dineen, Á. T., Truzzi, A., King, G., Zaadnoordijk, L., Harrison, K., D'Arcy, E.-L., White, J., Caldinelli, C., Holloway, T., Kravchenko, A., Diedrichsen, J., Tarrant, A., Byrne, A. T., Foran, A., Molloy, E. J. and Cusack, R. *in press*. Infants Have Rich Visual Categories in Ventrotemporal Cortex at 2 Months Old. *Nat Neuro*.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E. and Lindeløv, J. K. (2019) 'PsychoPy2: Experiments in behavior made easy', *Behav Res Methods,* 51(1), pp. 195-203.

Petras, K., Ten Oever, S., Jacobs, C. and Goffaux, V. (2019) 'Coarse-to-fine information integration in human vision', *Neuroimage,* 186, pp. 103-112.

Petro, L. S., Vizioli, L. and Muckli, L. (2014) 'Contributions of cortical feedback to sensory processing in primary visual cortex', *Front Psychol,* 5, pp. 1223.

Petroff, Z. J., Jayaraman, S., Smith, L. B., Candy, T. R. and Bonnen, K. (2025) 'The world through infant eyes: Evidence for the early emergence of the cardinal orientation bias', *Proc Natl Acad Sci U S A,* 122(16), pp. e2421277122.

Pollatou, A., Filippi, C. A., Aydin, E., Vaughn, K., Thompson, D., Korom, M., Dufford, A. J., Howell, B., Zöllei, L., Martino, A. D., Graham, A., Scheinost, D., Spann, M. N. and Group, F. N. (2022) 'An ode to fetal, infant, and toddler neuroimaging: Chronicling early clinical to research applications with MRI, and an introduction to an academic society connecting the field', *Dev Cogn Neurosci,* 54, pp. 101083.

Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L. and Petersen, S. E. (2012) 'Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion', *Neuroimage,* 59(3), pp. 2142-54.

Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L. and Petersen, S. E. (2014) 'Methods to detect, characterize, and remove motion artifact in resting state fMRI', *Neuroimage,* 84, pp. 320-41.

Power, J. D., Schlaggar, B. L. and Petersen, S. E. (2015) 'Recent progress and outstanding issues in motion correction in resting state fMRI', *Neuroimage,* 105, pp. 536-51.

Quinlan, P. T. (2003) *Connectionist models of development : developmental processes in real and artificial neural networks. Studies in developmental psychology* Hove, East Sussex ; New York: Psychology Press.

Quinn, P. C. and Eimas, P. D. (1996) 'Perceptual cues that permit categorical differentiation of animal species by infants', *J Exp Child Psychol,* 63(1), pp. 189-211.

Quinn, P. C., Eimas, P. D. and Rosenkrantz, S. L. (1993) 'Evidence for representations of perceptually similar natural categories by 3-month-old and 4-month-old infants', *Perception,* 22(4), pp. 463-475.

Quinn, P. C., Westerlund, A. and Nelson, C. A. (2006) 'Neural markers of categorization in 6-month-old infants', *Psychol Sci,* 17(1), pp. 59-66.

Raz, G. and Saxe, R. (2020) 'Learning in infancy is active, endogenously motivated, and depends on the prefrontal cortices', *Annual Review of Developmental Psychology,* 2(1), pp. 247-268.

Reh, R. K., Dias, B. G., Nelson, C. A., Kaufer, D., Werker, J. F., Kolb, B., Levine, J. D. and Hensch, T. K. (2020) 'Critical period regulation across multiple timescales', *Proc Natl Acad Sci U S A,* 117(38), pp. 23242-23251.

Reid, V. M., Dunn, K., Young, R. J., Amu, J., Donovan, T. and Reissland, N. (2017) 'The Human Fetus Preferentially Engages with Face-like Visual Stimuli', *Curr Biol,* 27(12), pp. 1825-1828.e3.

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A. and Ganguli, S. (2019) 'A deep learning framework for neuroscience', *Nature neuroscience,* 22(11), pp. 1761-1770.

Richardson, H., Lisandrelli, G., Riobueno-Naylor, A. and Saxe, R. (2018) 'Development of the social brain from age three to twelve years', *Nat Commun,* 9(1), pp. 1027.

Rosenblatt, F. (1958) 'The perceptron: A probabilistic model for information storage and organization in the brain', *Psychological Review,* 65(6), pp. 386–408.

Rosenke, M., Weiner, K. S., Barnett, M. A., Zilles, K., Amunts, K., Goebel, R. and Grill-Spector, K. (2018) 'A cross-validated cytoarchitectonic atlas of the human ventral visual stream', *Neuroimage,* 170, pp. 257-270.

Rumelhart, D. E., McClelland, J. L. and Group, P. D. P. R. (1986) *Parallel distributed processing, volume 1: Explorations in the microstructure of cognition: Foundations.* The MIT press.

Rust, N. C. and Movshon, J. A. (2005) 'In praise of artifice', *Nat Neurosci,* 8(12), pp. 1647-50.

Saffran, J. R. and Kirkham, N. Z. (2018) 'Infant Statistical Learning', *Annu Rev Psychol,* 69, pp. 181-203.

Santillo, M. F., O'Doherty, C., Dineen, Á. T., Truzzi, A., King, G., Zaadnoordijk, L., D'Arcy, E.-L., Holloway, T., White, J., Caldinelli, C., Kravchenko, A., Tarrant, A., Byrne, A. T., Foran, A., Molloy, E. J. and Cusack, R. *submitted*. The Origins of Memorability in Infancy.

Scholte, H. S., Smidi, J., Loke, J., Müller, N., Groen, I. I. A., van Gerven, M. A. J. and Smidi, J. 'Convolutional neural networks align early in training with neural representations'. *Cognitive Computational Neuroscience*, Amsterdam, The Netherlands.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J. and Geiger, F. (2018) 'Brain-score: Which artificial neural network for object recognition is most brain-like?', *BioRxiv*, pp. 407007.

Shultz, T. R. (2003) *Computational developmental psychology.* Cambridge, Mass.: MIT Press.

Siegel, J. S., Power, J. D., Dubis, J. W., Vogel, A. C., Church, J. A., Schlaggar, B. L. and Petersen, S. E. (2014) 'Statistical improvements in functional magnetic resonance imaging analyses produced by censoring high-motion data points', *Hum Brain Mapp,* 35(5), pp. 1981-96.

Simon, H. A. (1962) 'An information processing theory of intellectual development', *Monogr Soc Res Child Dev,* 27(2), pp. 150-61.

Sirois, S., Spratling, M., Thomas, M. S., Westermann, G., Mareschal, D. and Johnson, M. H. (2008) 'Précis of neuroconstructivism: how the brain constructs cognition', *Behav Brain Sci,* 31(3), pp. 321-31; discussion 331-56.

Siu, C. R. and Murphy, K. M. (2018) 'The development of human visual cortex and clinical implications', *Eye Brain,* 10, pp. 25-36.

Skelton, A. E., Maule, J. and Franklin, A. (2022) 'Infant color perception: Insight into perceptual development', *Child Dev Perspect,* 16(2), pp. 90-95.

Sloetjes, H. and Wittenburg, P. 'Annotation by category-ELAN and ISO DCR'. 2008.

Smith, L., Yu, C., Yoshida, H. and Fausey, C. M. (2015) 'Contributions of head-mounted cameras to studying the visual environments of infants and young children', *J Cogn Dev,* 16(3), pp. 407-419.

Smith, L. B. (2003) 'Learning to recognize objects', *Psychol Sci,* 14(3), pp. 244-50.

Smith, L. B., Jayaraman, S., Clerkin, E. and Yu, C. (2018) 'The developing infant creates a curriculum for statistical learning', *Trends in cognitive sciences,* 22(4), pp. 325-336.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., Bannister, P. R., De Luca, M., Drobnjak, I., Flitney, D. E., Niazy, R. K., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J. M. and Matthews, P. M. (2004) 'Advances in functional and structural MR image analysis and implementation as FSL', *Neuroimage,* 23 Suppl 1, pp. S208-19.

Solomon, J. A. and Pelli, D. G. (1994) 'The visual filter mediating letter identification', *Nature,* 369(6479), pp. 395-7.

Sowden, P. T. and Schyns, P. G. (2006) 'Channel surfing in the visual brain', *Trends Cogn Sci,* 10(12), pp. 538-45.

Spann, M. N., Wisnowski, J. L., Smyser, C. D., Howell, B., Dean, D. C., Group, H. P. I. S. Y. P. W. and Fetal, I. f., and Toddler Neuroimaging Group (FIT'NG) (2023) 'The Art, Science, and Secrets of Scanning Young Children', *Biol Psychiatry,* 93(10), pp. 858-860.

Spelke, E. S. (2002) 'Developmental neuroimaging: a developmental psychologist looks ahead', *Developmental Science,* 5(3), pp. 392-396.

St-Yves, G., Allen, E. J., Wu, Y., Kay, K. and Naselaris, T. (2023) 'Brain-optimized deep neural network models of human visual areas learn non-hierarchical representations', *Nat Commun,* 14(1), pp. 3329.

Storrs, K. R., Kietzmann, T. C., Walther, A., Mehrer, J. and Kriegeskorte, N. (2021) 'Diverse Deep Neural Networks All Predict Human Inferior Temporal Cortex Well, After Training and Fitting', *J Cogn Neurosci,* 33(10), pp. 2044-2064.

Subramanian, A., Sizikova, E., Majaj, N. and Pelli, D. (2023) 'Spatial-frequency channels, shape bias, and adversarial robustness', *Advances in neural information processing systems,* 36, pp. 4137-4149.

Suter, P. S., Suter, S., Roessler, J. S., Parker, K. L., Armstrong, C. A. and Powers, J. C. (1994) 'Spatial-frequency-tuned channels in early infancy: VEP evidence', *Vision Res,* 34(6), pp. 737-45.

Tansey, R., Graff, K., Rai, S., Merrikh, D., Godfrey, K. J., Vanderwal, T. and Bray, S. (2024) 'Development of human visual cortical function: A scoping review of task- and naturalistic-fMRI studies through the interactive specialization and maturational frameworks', *Neurosci Biobehav Rev,* 162, pp. 105729.

Tartaglini, A. R., Vong, W. K. and Lake, B. M. (2022) 'A developmentally-inspired examination of shape versus texture bias in machines', *arXiv preprint arXiv:2202.08340*.

Taylor, G., Hipp, D., Moser, A., Dickerson, K. and Gerhardstein, P. (2014) 'The development of contour processing: evidence from physiology and psychophysics', *Front Psychol,* 5, pp. 719.

Teller, D. Y. (1998) 'Spatial and temporal aspects of infant color vision', *Vision Res,* 38(21), pp. 3275-82.

Teller, D. Y. and Bornstein, M. H. (1987) 'Infant color vision and color perception', *Handbook of infant perception,* 1, pp. 185-236.

Teller, D. Y., McDonald, M. A., Preston, K., Sebris, S. L. and Dobson, V. (1986) 'Assessment of visual acuity in infants and children: the acuity card procedure', *Dev Med Child Neurol,* 28(6), pp. 779-89.

Teller, D. Y., Morse, R., Borton, R. and Regal, D. (1974) 'Visual acuity for vertical and diagonal gratings in human infants', *Vision Res,* 14(12), pp. 1433-9.

Torralba, A. and Oliva, A. (2003) 'Statistics of natural image categories', *Network,* 14(3), pp. 391-412.

Turk-Browne, N. B. and Aslin, R. N. (2024) 'Infant neuroscience: how to measure brain activity in the youngest minds', *Trends Neurosci,* 47(5), pp. 338-354.

Turkewitz, G. and Kenny, P. A. (1982) 'Limitations on input as a basis for neural organization and perceptual development: a preliminary theoretical statement', *Dev Psychobiol,* 15(4), pp. 357-68.

Uyar, F., Shomstein, S., Greenberg, A. S. and Behrmann, M. (2016) 'Retinotopic information interacts with category selectivity in human ventral cortex', *Neuropsychologia,* 92, pp. 90-106.

van Bergen, R. S. and Kriegeskorte, N. (2020) 'Going in circles is the way forward: the role of recurrence in visual inference', *Curr Opin Neurobiol,* 65, pp. 176-193.

van de Ven, G. M., Soures, N. and Kudithipudi, D. (2024) 'Continual learning and catastrophic forgetting', *arXiv preprint arXiv:2403.05175*.

van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., Yu, T. and contributors, s.-i. (2014) 'scikit-image: image processing in Python', *PeerJ,* 2, pp. e453.

Van Dijk, K. R., Sabuncu, M. R. and Buckner, R. L. (2012) 'The influence of head motion on intrinsic functional connectivity MRI', *Neuroimage,* 59(1), pp. 431-8.

Vanderwal, T., Kelly, C., Eilbott, J., Mayes, L. C. and Castellanos, F. X. (2015) 'Inscapes: A movie paradigm to improve compliance in functional magnetic resonance imaging', *Neuroimage,* 122, pp. 222-32.

Vaziri-Pashkam, M., Taylor, J. and Xu, Y. (2019) 'Spatial Frequency Tolerant Visual Object Representations in the Human Ventral and Dorsal Visual Processing Pathways', *J Cogn Neurosci,* 31(1), pp. 49-63.

Vogelsang, L., Gilad-Gutnick, S., Ehrenberg, E., Yonas, A., Diamond, S., Held, R. and Sinha, P. (2018) 'Potential downside of high initial visual acuity', *Proc Natl Acad Sci U S A,* 115(44), pp. 11333-11338.

Vogelsang, L., Vogelsang, M., Pipa, G., Diamond, S. and Sinha, P. (2024a) 'Butterfly effects in perceptual development: A review of the ‘adaptive initial degradation’hypothesis', *Developmental Review,* 71, pp. 101117.

Vogelsang, M., Vogelsang, L., Gupta, P., Gandhi, T. K., Shah, P., Swami, P., Gilad-Gutnick, S., Ben-Ami, S., Diamond, S., Ganesh, S. and Sinha, P. (2024b) 'Impact of early visual experience on later usage of color cues', *Science,* 384(6698), pp. 907-912.

Vogelsang, M., Vogelsang, L., Pipa, G., Diamond, S. and Sinha, P. (2025) 'Potential role of developmental experience in the emergence of the parvo-magno distinction', *Commun Biol,* 8(1), pp. 987.

Vogelsang, M., Vogelsang, L. and Sinha, P. (2026) 'Computationally Probing the Role of Time-Limited Neuronal Plasticity in Early Visual Development', *Developmental Science,* 29(1), pp. e70099.

Wandell, B. A., Dumoulin, S. O. and Brewer, A. A. (2007) 'Visual field maps in human cortex', *Neuron,* 56(2), pp. 366-83.

Wang, H., Wu, X., Huang, Z. and Xing, E. P. 'High-frequency component helps explain the generalization of convolutional neural networks'. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8684-8694.

Weiner, K. S., Barnett, M. A., Lorenz, S., Caspers, J., Stigliani, A., Amunts, K., Zilles, K., Fischl, B. and Grill-Spector, K. (2017) 'The Cytoarchitecture of Domain-specific Regions in Human High-level Visual Cortex', *Cereb Cortex,* 27(1), pp. 146-161.

Whitney, D. and Levi, D. M. (2011) 'Visual crowding: a fundamental limit on conscious perception and object recognition', *Trends Cogn Sci,* 15(4), pp. 160-8.

Wiesel, T. N. and Hubel, D. H. (1963) 'Single-cell responses in striate cortex of kittens deprived of vision in one eye', *J Neurophysiol,* 26, pp. 1003-17.

Wilcox, T. and Biondi, M. (2015) 'Object processing in the infant: lessons from neuroscience', *Trends Cogn Sci,* 19(7), pp. 406-13.

Wild, C. J., Linke, A. C., Zubiaurre-Elorza, L., Herzmann, C., Duffy, H., Han, V. K., Lee, D. S. C. and Cusack, R. (2017) 'Adult-like processing of naturalistic sounds in auditory cortex by 3-and 9-month old infants', *NeuroImage,* 157, pp. 623-634.

Xie, S., Hoehl, S., Moeskops, M., Kayhan, E., Kliesch, C., Turtleton, B., Köster, M. and Cichy, R. M. (2022) 'Visual category representations in the infant brain', *Curr Biol,* 32(24), pp. 5422-5432.e6.

Xu, Y. and Vaziri-Pashkam, M. (2021) 'Limits to visual representational correspondence between convolutional neural networks and the human brain', *Nat Commun,* 12(1), pp. 2065.

Xu, Z.-Q. J., Zhang, Y. and Luo, T. (2025) 'Overview frequency principle/spectral bias in deep learning', *Communications on Applied Mathematics and Computation,* 7(3), pp. 827-864.

Xu, Z.-Q. J., Zhang, Y. and Xiao, Y. 'Training behavior of deep neural network in frequency domain'. *International Conference on Neural Information Processing*, 2019: Springer, 264-274.

Yamins, D. L. and DiCarlo, J. J. (2016) 'Using goal-driven deep learning models to understand sensory cortex', *Nat Neurosci,* 19(3), pp. 356-65.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D. and DiCarlo, J. J. (2014) 'Performance-optimized hierarchical models predict neural responses in higher visual cortex', *Proceedings of the national academy of sciences,* 111(23), pp. 8619-8624.

Yates, T. S., Ellis, C. T. and Turk-Browne, N. B. (2021) 'The promise of awake behaving infant fMRI as a deep measure of cognition', *Current Opinion in Behavioral Sciences,* 40, pp. 5-11.

Yates, T. S., Fel, J., Choi, D., Trach, J. E., Behm, L., Ellis, C. T. and Turk-Browne, N. B. (2025) 'Hippocampal encoding of memories in human infants', *Science,* 387(6740), pp. 1316-1320.

Yates, T. S., Skalaban, L. J., Ellis, C. T., Bracher, A. J., Baldassano, C. and Turk-Browne, N. B. (2022) 'Neural event segmentation of continuous experience in human infants', *Proc Natl Acad Sci U S A,* 119(43), pp. e2200257119.

Yoshihara, S., Fukiage, T. and Nishida, S. (2023) 'Does training with blurred images bring convolutional neural networks closer to humans with respect to robust object recognition and internal representations?', *Front Psychol,* 14, pp. 1047694.

Zaadnoordijk, L., Besold, T. R. and Cusack, R. (2022) 'Lessons from infant learning for unsupervised machine learning', *Nature Machine Intelligence,* 4(6), pp. 510-520.

Zador, A. M. (2019) 'A critique of pure learning and what artificial neural networks can learn from animal brains', *Nat Commun,* 10(1), pp. 3770.

Zettersten, M., Foushee, R. and Goddu, M. K. (2025) ''Helpless' infants are active, goal-directed agents: response to Cusack et al', *Trends Cogn Sci,* 29(7), pp. 587-588.

Appendix 2

A2.1 Supplementary methods

A2.1.1 Calculating the maximum resolvable spatial frequency from display parameters

The infant's eyes were positioned approximately 27 cm (d) from the display. The visual angle subtended by the display was calculated separately for the horizontal and vertical dimensions using Equation A2.1, where $\ell = w$ for width and $\ell = h$ for height. At this viewing distance, the display subtended 56×31.5 degrees of visual angle (dva; $\theta_w \times \theta_h$). The image had a resolution of 640×360 pixels ($w_{\mathrm{px}} \mathrm{x}\, h_{\mathrm{px}}$). The sampling rate in pixels per degree (ppd) was calculated by dividing each pixel dimension (l_{px}) by the corresponding visual angle (θ_l). The maximum resolvable spatial frequency ($f_{\max}$) was then taken as half this value, corresponding to the Nyquist frequency (Equation A2.2) in cycles per degree (cpd). This yielded the same estimate in the horizontal and vertical dimensions: $f_{\max} = 5.7$cpd.

$$\boldsymbol{\theta_l = 2\arctan\left(\frac{l/2}{d}\right)} \qquad (\boldsymbol{A2.1})$$

$$\boldsymbol{f_{\max} = \frac{l_{\mathrm{px}}}{2\theta_l}} \qquad (\boldsymbol{A2.2})$$

A2.1.2 Simulated time course modelling to minimise perceptual confounds and maximise semantic spread

A2.1.2.A Verifying the contribution of perceptual confounds and semantic regressors

We first verified that each perceptual and semantic regressor could be reliably estimated within the fMRI design matrix. For each tag, a theoretical time course was constructed by convolving its event timings with a canonical hemodynamic response function. Synthetic blood-oxygen-level-dependent (BOLD-like) noise was then added (Cusack *et al.*, 2015) to generate responses for 1,000 independent simulated voxels. A first-level general linear model (GLM) was fit to each simulated dataset using Nilearn's *FirstLevelModel* (Contributors, 2022), and the resulting z-transformed t-statistics were extracted. For every tag, z-scores across simulated voxels were significantly greater than zero, indicating that the corresponding regressor could be reliably estimated within a standard GLM.

Having established this, we used simulated fMRI time-course modelling to derive stimulus subsets that minimised perceptual confounds, optimised statistical power for predefined contrasts, captured broad and representative semantic dimensions, and maximised discriminability among individual videos. The optimal subset produced by each modelling analysis informed the shortlist of videos that proceeded to adult piloting.

A2.1.2.B Optimising design efficiency for literature-defined semantic constructs

A greedy backwards elimination strategy, following (Linke and Cusack, 2015) was used to identify a subset of videos (n = 15) that maximised design efficiency, defined as

statistical power for detecting effects specified by a predefined set of contrasts (Supplementary Table A2.2; see Supplementary Methods section A2.1.2.D).

A2.1.2.C Optimising for stimulus distinctiveness

To identify a subset of videos (n = 15) that maximised stimulus distinctiveness, a similar greedy backwards elimination strategy was applied at the level of individual video regressors. This yielded a set of videos predicted to produce the most separable neural responses (see Supplementary Methods section A2.1.2.D).

A2.1.2.D Greedy backwards elimination strategy

The following greedy backwards elimination framework was used to optimise both design efficiency and stimulus distinctiveness. Starting from the full set of candidate videos, one video was iteratively removed at each step until the target subset size was reached. At each iteration, candidate subsets were generated by excluding each remaining video once, and a design matrix was constructed from the remaining *(n–1)* videos. Time courses were then obtained either by simulation or from empirically measured fMRI data, depending on the analysis.

Time-course simulation. For each candidate subset, fMRI response time courses were simulated for 1,000 independent voxels by adding synthetic BOLD-like noise to appropriately weighted regressors from the candidate design matrix.

GLM Fitting. A first-level GLM was then fit to each time course, and an analysis-specific selection criterion was computed and z-transformed to provide a common scale.

Selection criterion for design efficiency. Z-transformed *t*-statistics were computed for each predefined contrast, and mean z(*t*) values were averaged across simulated voxels and then across contrasts to yield a single summary score for each candidate subset.

Selection criterion for stimulus distinctiveness. Each video was modelled as a separate regressor within the design matrix. An omnibus F-test was computed to assess whether any individual video regressor explained variance beyond the other regressors; resulting F-statistics were z-transformed, and mean z(F) values were averaged across simulated voxels to yield a single separability score for each candidate subset.

At each iteration, the video whose exclusion yielded the highest mean z-transformed selection-criterion value, that is, the greatest improvement in the criterion, was permanently removed. This leave-one-out procedure was then repeated on the reduced set until the target subset size was reached. To assess robustness, the full optimisation was repeated 100 times using independent samples of synthetic BOLD-like noise, and videos most frequently retained across runs were selected as the final optimised stimulus set.

These analyses were used in conjunction with manual selection to reduce the number of videos proceeding to adult piloting (n = 16).

A2.1.3 Adult piloting

A series of adult pilot studies was conducted to inform further reduction of the stimulus set.

In the first adult pilot fMRI study, adult participants (n = 16) viewed two runs of the video paradigm, each containing one repetition per video, and three runs of the pictures paradigm, each containing three repetitions per picture. Group-level responses to the stimuli were characterised using representational similarity analysis (RSA) in two broad ventral visual stream ROIs: early visual cortex and ventral temporal cortex (Kriegeskorte, Mur and Bandettini, 2008). For each paradigm, across-subject representational similarity matrices (RSMs) were computed to characterise group-level similarities in response patterns across contexts, with each stimulus modelled as a context-specific event. Response patterns elicited by each video were also compared with those elicited by the corresponding picture stimuli.

A greedy elimination strategy similar to that described above was then applied using adult brain responses in place of simulated data to inform manual selection of the six contexts that proceeded to the second adult pilot.

For the second adult pilot, the corresponding pictures from the first pilot were retained for each shortlisted video. Additional object categories were then added, increasing the total number of object categories per context from two to four. This provided a broader initial selection of objects to help maximise the power of the final design. As before,

each object category was represented by three exemplars, giving a total of 72 object images.

In the second adult pilot, participants (n = 17) viewed two runs of the video paradigm, each containing three repetitions per video, and three runs of the pictures paradigm, each containing three repetitions per picture. As before, RSMs were computed for the reduced shortlist of videos, the revised set of pictures, and the relationship between pictures and videos across contexts. An additional RSM was computed across individual objects to assess whether they clustered according to context. To quantify the reliability of responses across participants, an inter-subject correlation (ISC) analysis was also performed on the video stimuli (see section A2.1.3.A). One video did not elicit a reliable ISC and was therefore replaced in the final shortlist by an alternative from adult pilot 1.

A2.1.3.A Inter-subject correlation analysis

For each region of interest (ROI) and each participant, an average time course was extracted for each video by averaging across repetitions and runs. Inter-subject correlation (ISC) was then calculated by correlating each participant's time course for a given video with the average time course of the remaining participants for the same video. A null distribution was generated by repeatedly selecting a random participant and video, with replacement, and correlating that participant's time course with the average time course of the remaining participants for a different, randomly selected video. For each video, the significance of the ISC was determined by comparing the

distribution of ISC values across participants, for a given video and ROI, to the null distribution using a Mann-Whitney U test.

A2.1.3.B Final stimulus shortlisting based on adult piloting

To increase the power of the paradigm for infants, the number of picture stimuli was reduced from four object categories per context to two, mirroring the reduction applied to the videos. This reduction was informed by hierarchical clustering of the picture RSM. Stimuli that clustered with their associated context were preferentially selected. The final object set was balanced across animate, inanimate small, and inanimate large classifications.

A2.1.3.C Ethics and data from piloting

Ethical approval for the adult MRI studies was provided by the School of Psychology Research Ethics Committee at Trinity College Dublin (SPREC102018-19, approved 18 December 2018). Further details of the adult fMRI acquisition are provided in related work reporting analyses of these datasets. A subset of the picture data acquired in adult pilot 2 became the adult comparison cohort for the pictures task. Further details of the acquisition for adult pilot 2 are therefore available in O'Doherty et al. (2026). Due to the substitution in the final video shortlist, adult pilot 1 became the adult comparison cohort for the videos task. Further details of the acquisition for adult pilot 1 will be provided in Santillo *et al.* (submitted). Acquisition details and results for the video task and the objects-and-contexts task will be reported in subsequent publications.

A2.1.4 Details of paradigms

A2.1.5 Video order

Six predefined orders (A–F) were used to maintain consistent stimulus timing across subgroups of participants, enabling inter-subject correlation analyses. The orders were created pseudo-randomly with each video constrained to appear once in every serial position across the six orders. To allow for within-subject MVPA, each run of the videos paradigm consisted of two repetitions of the full set of videos. To eliminate temporal confounds in within-subject MVPA, a different order was used for each repetition, making sure that particular conditions were not consistently closer in time to other conditions, on average across participants. For each participant, the same pair of orders was used across all runs. The first order was chosen randomly. As each infant would participate in a variable number of runs, the second order was selected adaptively. Excluding the randomly chosen order, the order with the lowest cumulative number of runs within a given age cohort was identified and assigned as the second order. Thus, each infant viewed two orders, and the overall frequency distribution of the orders remained approximately balanced within each cohort. To direct infant attention to the display, prior to the onset of the videos an attention getter, which consisted of a rotating object accompanied by a noise such as a bell, was used both at the beginning of the paradigm and between orders. The attention getter was chosen randomly from a set of attention getters with proven efficiency at capturing infant attention (Bergmann *et al.*, 2020). To maximise engagement, the videos paradigm included the accompanying audio of speech or music.

A2.1.6 Pictures

For the pictures task, objects were presented in a pseudo-random order, comprising subblocks that contained one example from each category. Unlike the videos task, the presentation order varied across runs for each participant. Each object was presented for 3 seconds, interleaved with a white fixation cross. The fixation was presented for a jittered duration, ranging from 0.5–1.5 s with the hope of allowing a stimulus minus baseline analysis. Greater jitter would have been preferable but it was unclear whether a longer interval would be tolerated by the infants. As with the videos, each run of the pictures paradigm consisted of two repetitions of the full set of objects, with an attention getter used prior to each repetition. The audio accompanying the pictures consisted of nursery rhymes that were unrelated to the visual stimuli.

A2.2 Supplementary results

A2.2.1 Infants included in analysis of attentive state

Due to a combination of technical and hardware issues, and manual error, no facial recording was acquired for 20 of the runs from these infants (5%) at 2-months, and 8 (5%) at 9-months. Of the acquired runs, at 2-months 37 (10%) were partial recordings, that is, they did not encompass the full fMRI run, due to either a delay in starting the recording or premature termination. At 9-months, 47 (34%) runs were partial recordings. No recordings were acquired from six 2-month-olds (20 runs) and eight 9-month-olds (15 runs). The recordings from the remaining 31 2-month-olds and 11 9-month-olds had timing errors which precluded precise alignment with the fMRI acquisition and were therefore excluded from the present analysis. Despite these exclusions, a substantial,

randomly sampled, corpus of in scan facial recordings were tagged for attentive state. In-scan facial recordings were available for tagging from 93 2-month-olds (369 runs; 233 video runs, 136 picture runs) and 46 9-month-olds (138 runs; 80 video runs, 58 picture runs). This amounted to approximately 30 hours of awake scan time at 2-months, and 11 hours at 9-months.

A2.3 Supplementary figures

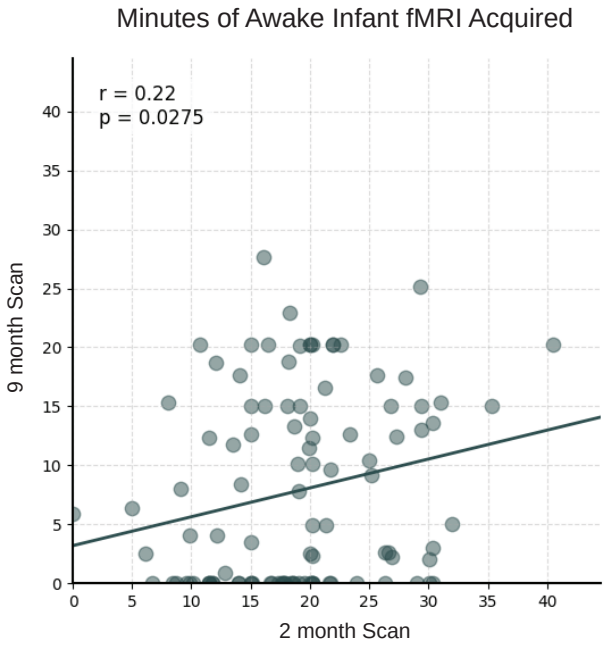

Figure A2.1 The amount of awake functional MRI data acquired at the 2-month scan was weakly but significantly predictive of the amount of data that would be acquired at the 9-month scan.

A2.4 Supplementary Tables

Table A2.1 Semantic tag definitions

Semantic Tag	**Definition**
Faces	When a face(s) is present and recognisable in the scene. Any object with eyes is considered a face.
Body parts	When a body part(s) is present and recognisable in the scene.
Scenes	Defined by exclusion: this tag applies when there is not a close-up of a face or object and a '*scene'* is visible.
Tools	When a tool(s) is recognisable in the scene.
Tool use	When a tool is salient, and an agent is holding the tool and using the tool.
Animate	Any character or object that moves on its own.
Inanimate big and inanimate small:	Refers to properties of objects, not the absolute size of the object on the screen - bath would be inanimate big, rubber duck would be inanimate small even if zoomed in. Exceptions: Clothes being worn are not tagged. Clothes not being worn would receive an inanimate small tag.
Social	Any time two or more characters are present in a scene and are interacting.
Non-social	Any time fewer than two characters are present in a scene or more than two characters are present and are not interacting.

Biological	All living things, including people, animals, plants etc. Excludes processed edible objects.
Non-biological	Any non-living object
Nature	The scene is set in a natural landscape (tag for duration of the scene). Precludes civilisation.
Civilisation	The scene is set in an urban landscape (tag for duration of the scene). Precludes nature.
Scene change	Any time the scene changes.
Camera cut	Any time the camera angle changes abruptly (does not include smooth camera motion across the frame) includes scene changes.
Biological motion	Any object that is moving in a biological manner (and that is salient in that a viewer would likely notice this object moving while focusing on the image as a whole).
Salient near towards	When a salient object in the foreground of a scene moves towards the viewer.
Salient near away	When a salient object in the foreground of a scene moves away from the viewer.

Supplementary Table A2.2. Literature-based contrasts of interest

Contrast	Contrast Weight(s)	Reference
Faces	+1	(Kanwisher, McDermott and Chun, 1997)
Body parts	+1	(Downing et al., 2001)
Scenes	+1	(Epstein and Kanwisher, 1998)
Tools	+1	(Chao, Haxby and Martin, 1999)
Animate, inanimate (small), inanimate (big)	+1, −0.5, −0.5	(Konkle and Caramazza, 2013)
Biological motion	+1	
Nature, civilisation	+1, −1	(Kravitz, Peng and Baker, 2011)
Social, non-social	+1, −1	(Huth et al., 2012)
Biological, non-biological	+1, −1	(Huth et al., 2012)
Far, open, near, closed	+1/3, +1/3, −1/3 ,−1/3	(Kravitz, Peng and Baker, 2011)

Appendix 3

A3.1 Supplementary methods

A3.1.1 Participant recruitment

Participants were recruited via a convenience sample from The Coombe Hospital and the Rotunda Hospital into one of two cohorts, the control group or the neonatal intensive care unit (NICU) group. Infants in the control group were born at term (between 38-42 weeks of gestation) and were without diagnosis of brain injury or congenital abnormalities. Infants in the NICU group were preterm-born infants that had spent time in the NICU. Infants were recruited into the control group by a researcher from the Foundations of Cognition (FOUNDCOG) project team. Researchers attended the postnatal wards of the participating maternity hospitals and approached caregivers to ask whether they would be interested in hearing about the study, and to provide them an information leaflet. Infants were similarly recruited into the NICU group by a Neonatal Specialist Registrar who approached caregivers in the NICU of the same hospitals. Following a short discussion, if caregivers were interested, they provided the researcher with contact information and the due date of their infant. The due date was used to estimate the corrected gestational age (CGA) of the infant. Interested caregivers were contacted when their infant was 1-month CGA. If they wished to participate, the infant was deemed eligible, and no contraindications to MRI were identified, they were invited to attend a scanning session when their infant was 2 months CGA. At the session, informed consent was obtained from caregivers on behalf of their infant and further MRI safety screening was conducted.

A3.1.2 Functional MRI setup and acquisition

Participants were scanned in a Siemens 3T MAGNETOM Prisma located at the Trinity College Institute of Neuroscience, on the main campus of Trinity College Dublin. Each scan was attended by a radiographer and a clinician (doctor or nurse) who had completed Advanced Paediatric Life-Support training, in addition to at least two further staff members who were typically a research scientist and a research assistant. As described in prior work (O'Doherty *et al.*, *in press*), scan setup was designed to maximise comfort and minimise motion, inspired by other awake infant scanning groups (Ellis *et al.*, 2020; Deen *et al.*, 2017). Infants were placed on a bead-filled vacuum pillow on the scan bed. Infants wore Optoacoustics OptoActive II active noise cancelling headphones, held in place by inflatable cushions which lined the posterior head coil. This ensured that the headphones formed a seal, providing hearing protection and allowing audio stimulus delivery. Engaging visual stimuli were projected on the bore of the scanner, as described in (Chapter 3, Methods section 3.5.2). A surface coil was used in lieu of the anterior head coil to ensure infants could see stimuli on the bore. Infants could use a pacifier if they desired and did so approximately half of the time. A team member remained beside the scanner with the infant during the scan, with simultaneous real-time monitoring of infant state from the control room using an MRC Systems in-bore camera recording of the face of the infant. Infants were scanned using the posterior array of a Siemens 64-channel head and neck coil. At 2 months, a 4-channel flexi-coil was also placed on a shelf above the forehead of the infant to capture signal from the frontal lobes. Infant fMRI was acquired with a multiband EPI sequence, with an in-plane acceleration factor of 4, repetition time (TR) of 610 ms, echo time (TE) of 32 ms, flip angle (FA) of 40^0, anterior-to-posterior phase-encoding direction, 3 mm isotropic voxels, field of view (FOV) of 192 × 192 mm, and 64 × 64 × 36 voxels per volume. Adults were scanned using the full 64-channel head coil. Adult fMRI was

similarly acquired with a multiband EPI sequence with an in-plane acceleration factor of 4 and 3 mm isotropic voxels. Other parameters varied slightly: TR = 656 ms, TE = 30 ms, FA = 50 ^{0}, left-to-right phase-encoding direction, FOV = 224 × 245 mm, 40 slices per volume.

A3.1.3 Functional MRI analysis and preprocessing

Data preprocessing and analysis of infant data was performed using a custom pipeline developed by the FOUNDCOG team specifically for awake infant fMRI using Nipype (Gorgolewski *et al.*, 2011). This pipeline is openly available at https://github.com/ClionaOD/foundcog_analysis. Briefly, for each functional run, volumes (EPI images) were motion corrected using FSL MCFLIRT (Jenkinson *et al.*, 2002), and their framewise displacement (FD) estimated. For each subject, susceptibility-induced distortions were estimated from field maps acquired with opposite phase encoding directions (AP/PA) and used to correct the EPI images using FSL topup (Andersson, Skare and Ashburner, 2003; Smith *et al.*, 2004). The functional runs were then co-registered using FLIRT (Jenkinson and Smith, 2001; Jenkinson *et al.*, 2002) with an affine transformation, with the degrees of freedom (6, 9 or 12 – typically 12) chosen for each subject based on manual inspection. Each subject's data was then normalised to common space using an NIHPD infant template (2-5 month or 8-11 month; Fonov *et al.*, 2011) by application of a further affine transform. Adult data was preprocessed using fMRIPrep (Esteban *et al.*, 2019) and was normalised to the MNI152 template (Mazziotta *et al.*, 2001).

For each functional run, a volumetric whole-brain general linear model (GLM) was fit using Nilearn (version 0.9.2; Contributors, 2022). A design matrix was constructed by modelling

each of the 36 objects as a regressor and convolving with a canonical hemodynamic response function. Nuisance regressors included the 6 motion parameters, a linear trend regressor and a cosine drift model with a high pass filter of 0.01 Hz. A motion threshold of 1.5 mm FD was chosen, with volumes exceeding this threshold censored (excluded) from analysis (Power *et al.*, 2014). Runs were excluded from analysis if more than half of the volumes exceeded this threshold, or if the researcher attending the scan deemed the run not to be of sufficient quality based on observation. Included timepoints had a median FD of 0.308 mm (IQR = [0.163-0.637]) at 2 months and 0.267 mm (IQR = [0.161-0.486]) at 9 months. A final sample of 101 2-month-olds, 44 9-month-olds and 17 adults contributed to analysis and subsequent group-level visual object representations.

A3.1.4 Split-half noise ceiling calculation

To estimate the noise ceiling for each ROI, split-half reliability was obtained by dividing the data from each cohort to create two group-level representational similarity matrices and calculating their correlation. The Spearman-Brown prophecy formula was then applied to adjust for the reduction in sample size that occurs when splitting the data (see Table A3.1).

A3.1.5 Deep neural network training implementation details

ResNet-50 (He *et al.*, 2015) was trained on the ImageNet training dataset (Deng *et al.*, 2009) using supervised learning. Networks were trained for 60 epochs with a batch size of 256 using a standard hyperparameter configuration. Training used stochastic gradient descent with momentum of 0.9, a learning rate scheduler with an initial learning rate of 0.1 which decayed by a factor of 10 every 30 epochs, and a standard data augmentation procedure. Images were

randomly cropped to 224 by 224 pixels, blurred when applicable using the torchvision GaussianBlur transform, and were randomly flipped horizontally, following which the image was converted to a tensor and normalised to the ImageNet mean [0.485, 0.456, 0.406] and standard deviation [0.229, 0.224, 0.225]. Training was implemented in Python 3.7.7 with PyTorch 1.7.1 and CUDA 10.1, on a Lambda Quad GPU workstation with 28 CPU cores, 128 GB of system RAM, equipped with 4 NVIDIA RTX 6000 GPUs, each with 24 GB of memory.

A3.1.6 Characterising the spatial frequency of visual features extracted by blur-trained DNNs

For models trained on each level of blur, we identified a set of image patches from the model training dataset that led to high levels of activation in individual neurons from a chosen layer. Patches were identified from sharp images, which contain all spatial frequencies, thus serving as a broadband probe to test whether blurry trained models preferentially responded to image patches containing lower spatial frequencies relative to sharp trained models.

Patches were identified using the two-step process of Görgün and colleagues (https://github.com/adagorgun/VITAL; Görgün, Schiele and Fischer, 2025). We characterised the spatial frequency preference of an intermediate model layer (layer 31 – the output of residual stage 3 block 2, with 1024 neurons), analysing responses from a random subset of 64 neurons. For each of these neurons, we identified the 10,000 images that led to the highest levels of activation. From each image, 16 patches of size 64 × 64 px were extracted. The 16 patches from distinct images that led to the highest levels of activation for a given neuron were chosen as the preferred image patches for that neuron, yielding 1024 patches (16 patches for 64 neurons). As two models were trained per blur level, this process was repeated

for each, yielding a total sample of 2048 preferred patches for the chosen layer trained on a given level of blur.

For each level of blur, the power spectrum was calculated and averaged across patches (Figure A3.3A). This was compared to the ImageNet baseline, which was calculated by taking a random sample of 16 patches from 500 images, giving a total of 8,000 patches. To compare the model-wise spectra to the ImageNet baseline, the ratio of each power spectrum relative to baseline was calculated for each blur level (Figure A3.3B).

A3.2 Supplementary tables

Table A3.1 Split-half noise ceilings for regions of interest along the ventral visual stream.

ROI	2 months	9 months	Adult
V1	0.9422	0.9383	0.9516
V2dv	0.9010	0.9077	0.9399
V3v	0.9463	0.9432	0.9458
hV4	0.8495	0.8569	0.8488
VO1	0.6954	0.7460	0.6618
FG4	0.3211	0.6016	0.7263

A3.3 Supplementary figures

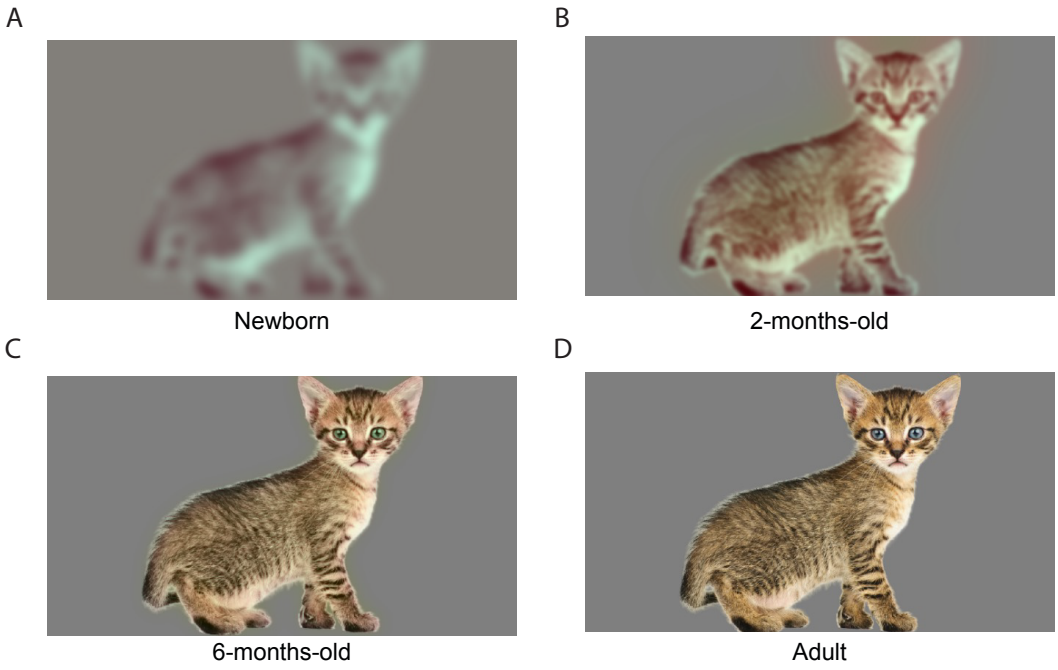

Figure A3.1 Sample object as seen by an infant viewing the FOUNDCOG experimental display over the first postnatal year.

Visual acuity and colour vision develop rapidly over the first postnatal year (Brown and Yamamoto, 1986; Courage and Adams, 1990; Brown, 1990; Adams, Courage and Mercer, 1991; Teller, 1998; Adams and Courage, 2002); consequently, the test-time appearance of visual stimuli will also change. To estimate the appearance of our stimuli at 2-month corrected gestational age, we used the TinyEyes simulator (https://github.com/wadelab/VischeckTinyeyes), which simulated infant vision accounting for display parameters at test-time. Given our experimental paradigm and display parameters, which consisted of large objects (spanning up to 29 cm in width) lacking in high spatial frequency details (maximum 5.7 cycles per degree of visual angle), presented close to the face of the infant (27 cm), the identities of the presented objects were clearly identifiable at 2 months (B). The estimated appearance of the same stimulus to a newborn (A), a 6-month-old (C) and an adult (D) are shown for comparison.

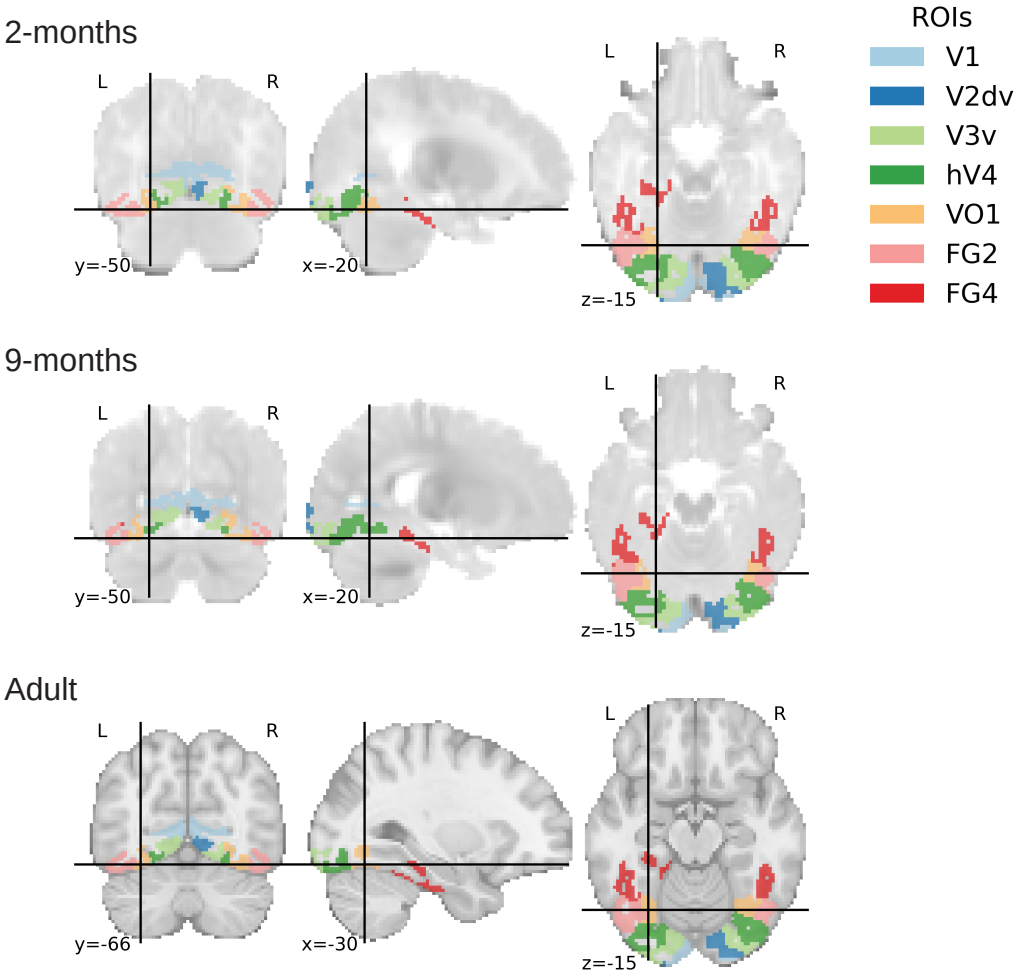

Figure A3.2 Ventral visual stream regions of interest for which tuning to visual object features was characterised.

Regions of interest (ROIs) were defined using cytoarchitectonic parcellations from the Jülich atlas (Amunts *et al.*, 2020a; Amunts *et al.*, 2020b), which have been previously cross-validated as overlapping with functional regions of the ventral visual stream in adults, as described in (Chapter 3, Methods section 3.5.5.A). ROIs were normalised to age-appropriate templates and are visualised for each age group on the corresponding template. An NIHPD template was used for infants (Fonov *et al.*, 2011) and the MNI template was used for adults (Mazziotta *et al.*, 2001).

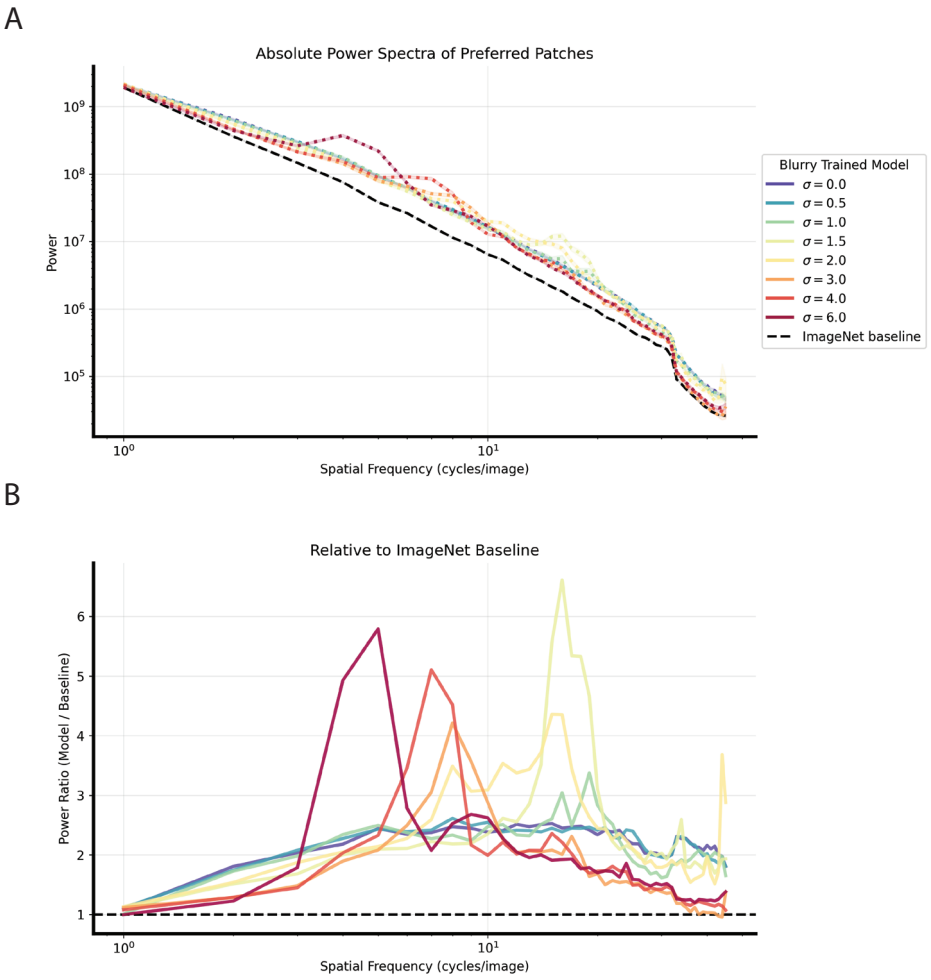

Figure A3.3 Blurrier trained DNNs preferred coarser object features.

For ResNet-50 models trained on each level of blur (σ = 0–6), we identified a set of image patches (n = 2048) from the standard, sharp model training dataset (σ = 0), that led to high levels of activation in neurons in an intermediate model layer (layer 31). The average power spectrum of the preferred patches for each blur level was calculated and is depicted in (A), with the error band representing 1000 bootstrap resamples of the data. The baseline power spectrum was estimated from a random sample of patches (n = 8000) from the same dataset (dashed black line). As expected for natural images, the baseline power spectrum showed an approximately power-law decay with spatial frequency (typically found to be close to $1/f^2$), as evidenced by a near-linear trend on the log–log plot (Field, 1987; Torralba and Oliva, 2003). The ratio of the power spectrum for the blurry trained models relative to this baseline is depicted in (B).

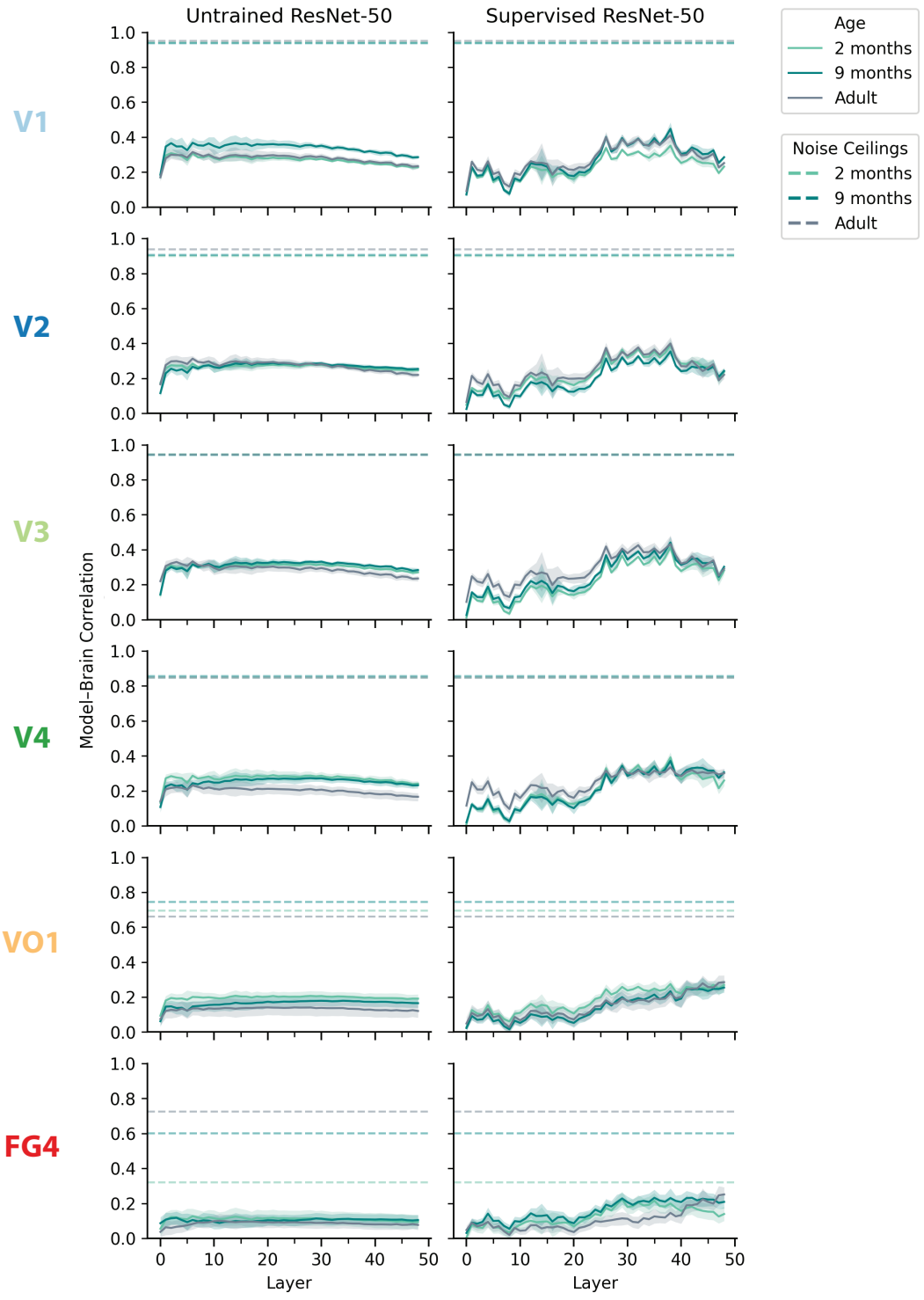

Figure A3.4 ResNet-50 captures developing object responses along the ventral visual stream.

Layer-wise correspondence between developing visual responses along the ventral visual stream (VVS) and representations from ResNet-50. Panels on the left show

model–brain correspondence for untrained models, whereas panels on the right show correspondence after supervised training on ImageNet. For each age group and region, line plots depict the median model–brain correlation across 1000 bootstrap resamples, with error bands showing the 95% confidence interval. Two iterations of each model were used, corresponding to two random initialisations; 500 bootstraps were compared to each iteration, giving 1000 bootstraps per condition. Dashed liindicate the age-specific split-half noise ceiling for each region.

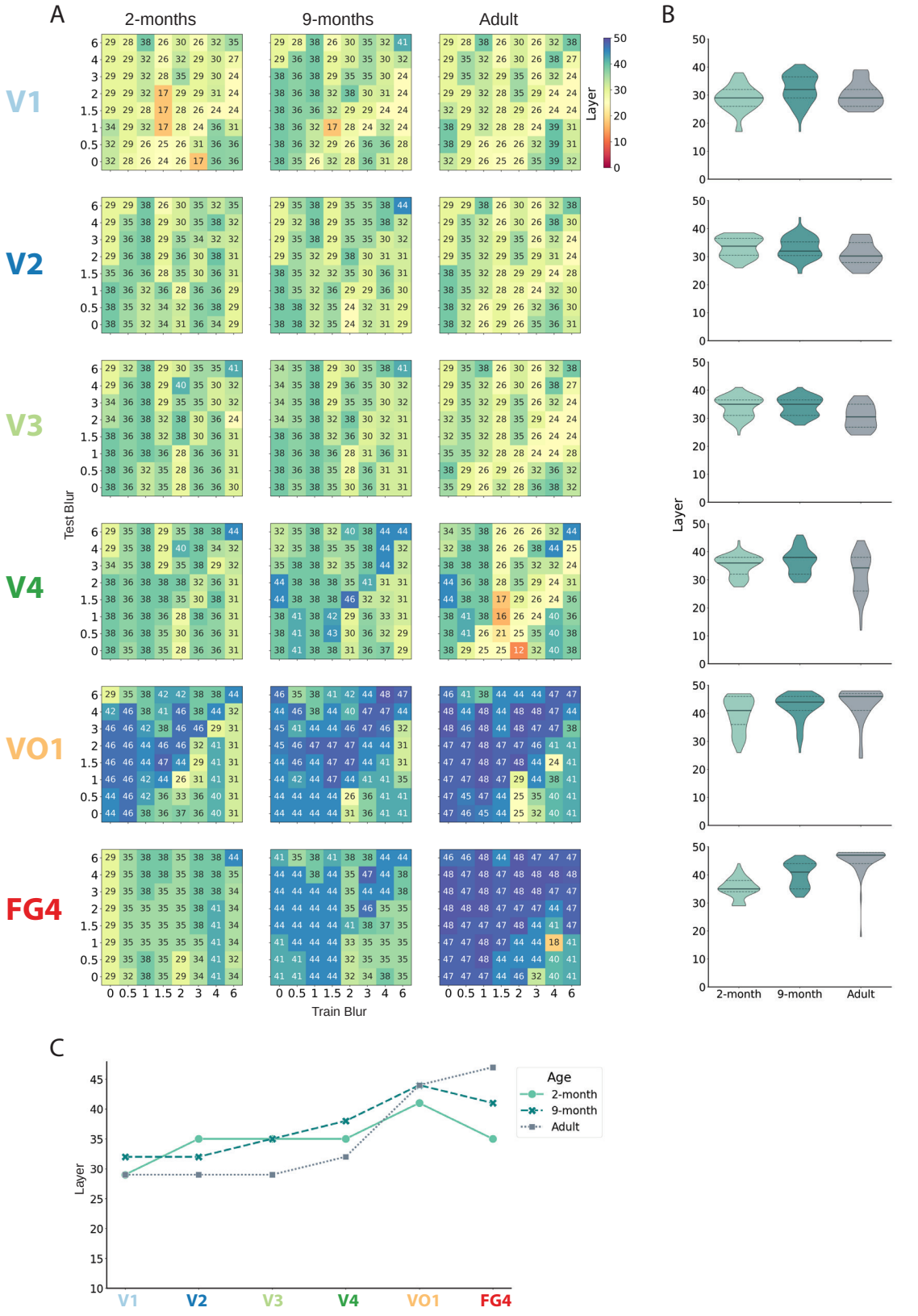

Figure A3.5 Model layers that best capture multiscale object feature tuning along the developing ventral visual stream.

Early regions (V1–V4) were best modelled by intermediate model layers, while VO1 was better captured by later layers across development. A later region, FG4, showed a divergence in the layers that best capture its visual representations across development: 2-month-olds were better represented by earlier layers, suggesting that this region encodes features of lower complexity in young infants. For a given ROI and age, for each train-test combination, the number of the layer that best captured the visual representation for each bootstrap of the data was recorded for 1000 bootstrap resamples, producing a distribution of layer numbers for each region and age. The median of the distribution was considered the layer that best characterised a given train-test combination; the median layer number for each train-test combination is shown in (A). For each age and ROI, the values plotted in (A) are depicted as a distribution in (B). The thick black line in the centre of the violin plot depicts the median value, the dashed black lines depict the inter-quartile range. The median value was extracted from each distribution and is plotted in (C), summarising the layer that best characterised a given region across all train-test combinations.

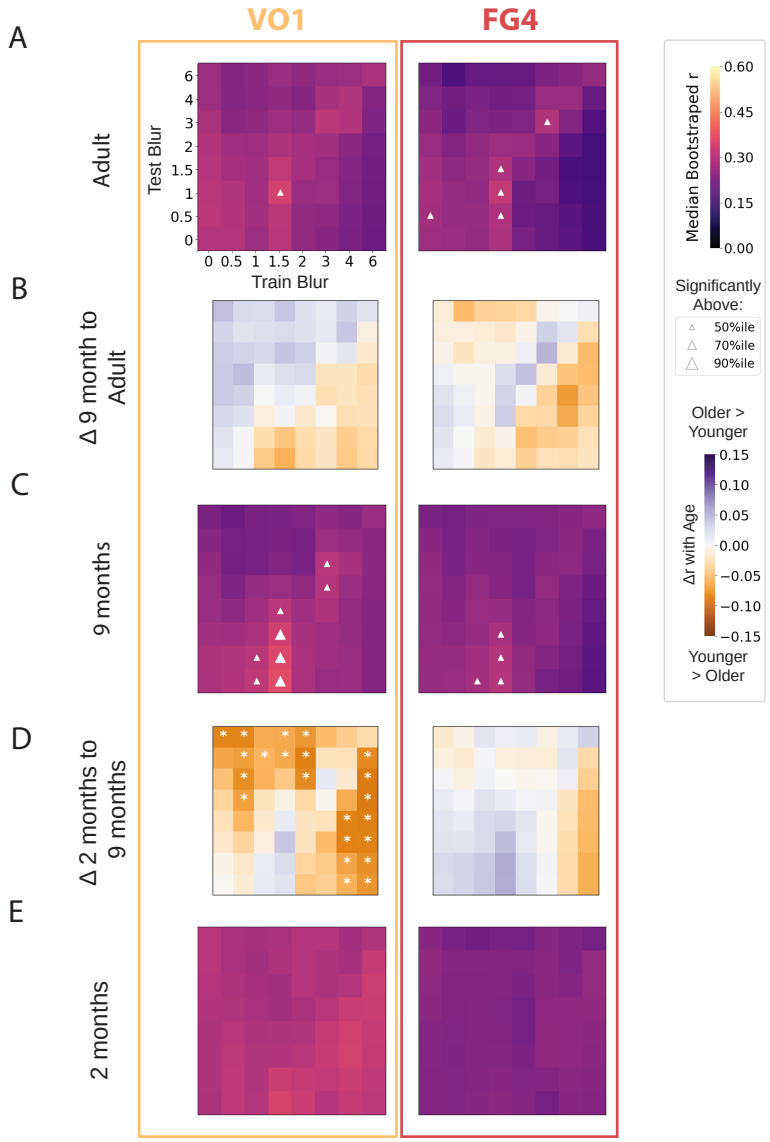

Figure A3.6 Multiscale object feature tuning for later visual regions.

Multiscale tuning maps for later ventral stream regions — VO1, an intermediate visual region, and FG4, a region in category-selective cortex — are shown for adults (A), 9-month-olds (C), and 2-month-olds (E). Tuning maps were calculated as in (Chapter 3, Figure 3.2A,C,E). Developmental differences in tuning patterns between 2 months and 9 months (D), and 9 months and adults (B) were calculated by subtracting the tuning map of the younger age from that of the older age, as described for (Chapter 3, Figure 3.2B and 2D).

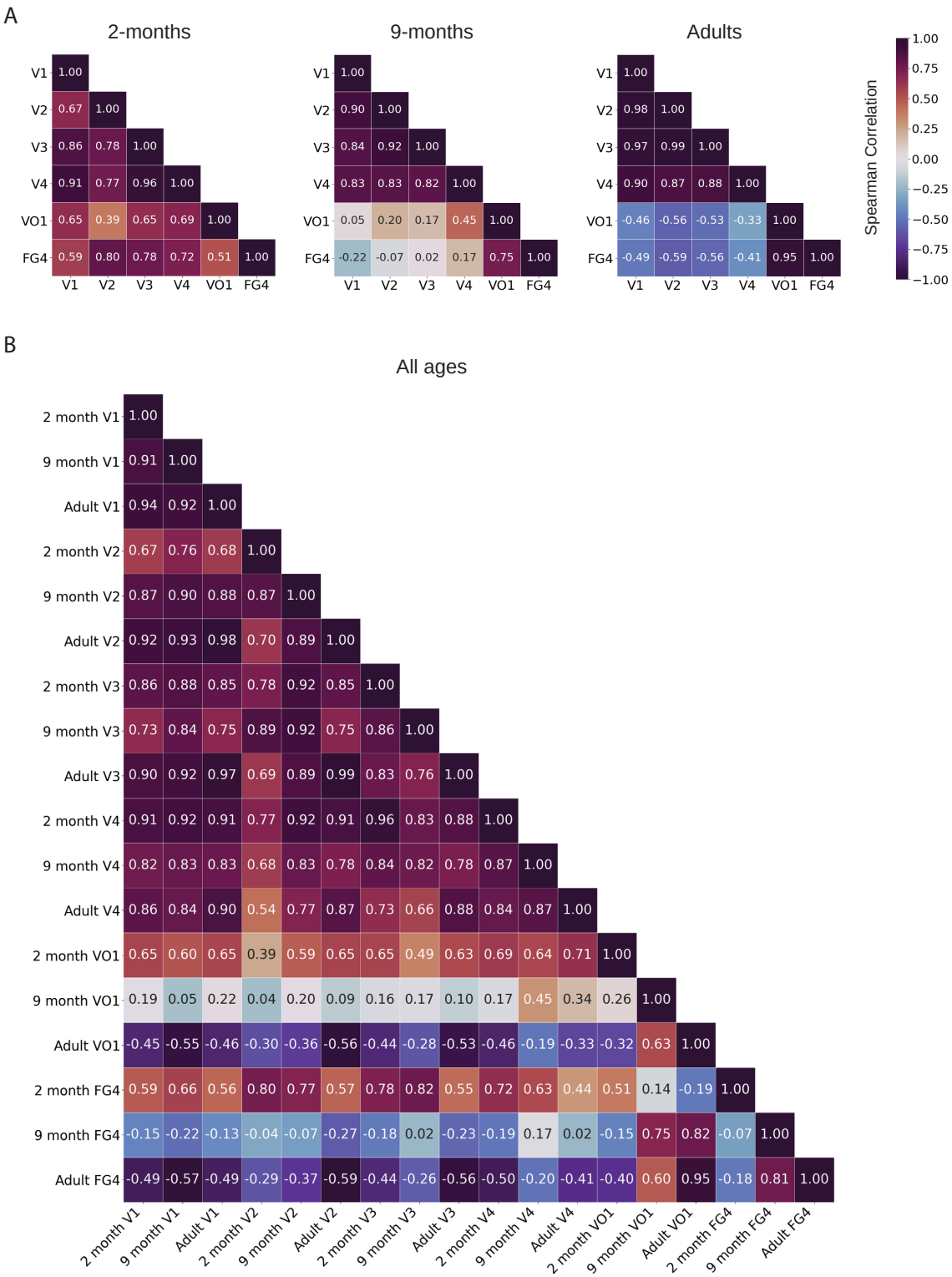

Figure A3.7 Similarity of multiscale tuning maps along the ventral visual stream across development.

Spearman correlation between tuning maps within each age group (A). At 2 months, the features encoded by VO1 and FG4 were found to be more similar to the early visual cortex (EVC) than they were at older ages. By 9 months, the pattern of tuning in these regions was distinct from that of the EVC and by adulthood the tuning patterns in early and later visual regions are negatively correlated. Spearman correlation between the tuning maps along the ventral visual stream between age groups highlights the similarity in the tuning pattern throughout the EVC across development (B).

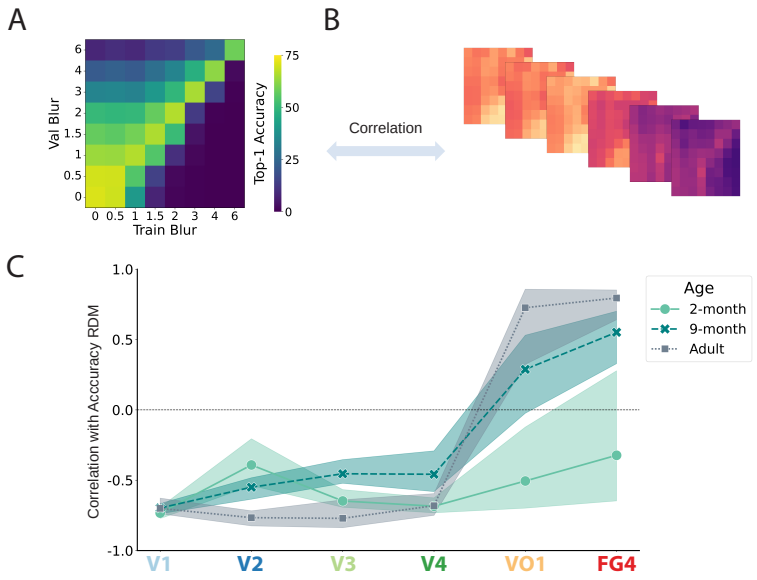

Figure A3.8 Correlation between tuning maps and blur-dependent model classification accuracy.

Classification accuracy was calculated for each train-test combination of blur by calculating the top 1 classification accuracy of each model in response to the ImageNet validation images (val) and averaging across model repetitions (A; as described in Chapter 3, Methods section 3.5.6.D). The matrix of blur-dependent model accuracy was correlated with tuning maps along the VVS (B). High levels of similarity would indicate that a region at a given age was correlated well with models with high classification accuracy. Median bootstrapped correlation with 95% confidence interval between tuning maps and the blur-dependent accuracy map along the VVS across development (C). The early visual cortex was inversely correlated with the train-test combinations that produced the highest model accuracy across development, whereas later VVS regions became progressively more correlated with model accuracy across development.

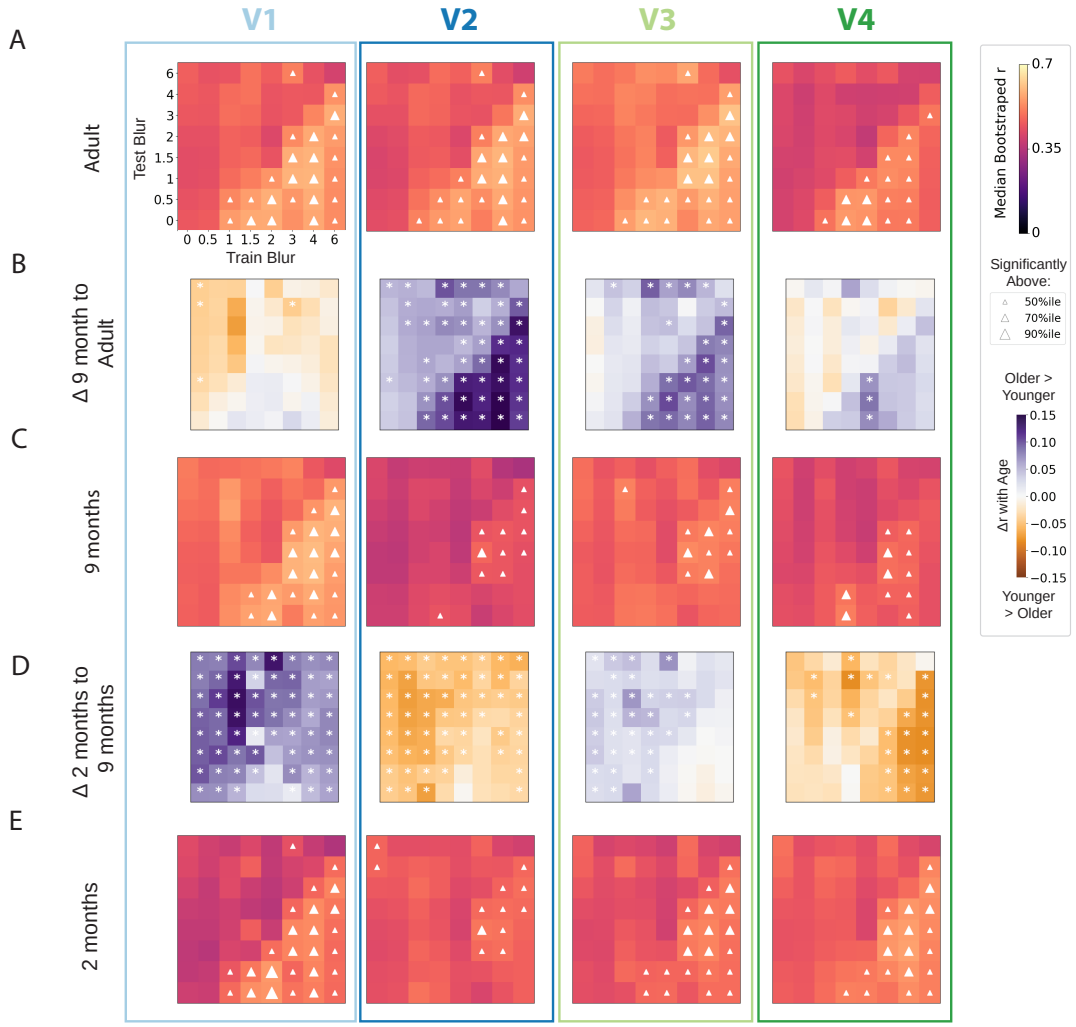

Figure A3.9 Multiscale object feature tuning for early visual regions scaled by the noise ceiling.

Multiscale tuning maps for the early visual cortex of adults (A), 9-month-olds (C) and 2-month-olds (E) scaled by the appropriate noise ceiling (as defined in Table A3.1). Tuning maps were calculated and are depicted as described for Chapter 3, Figure 3.2A,C,E. Developmental differences in the scaled tuning maps between 2 months and 9 months (D), and 9 months and adulthood (B), were calculated by subtracting the tuning map of the younger age group from that of the older age group, as described for Chapter 3, Figure 3.2B and D.

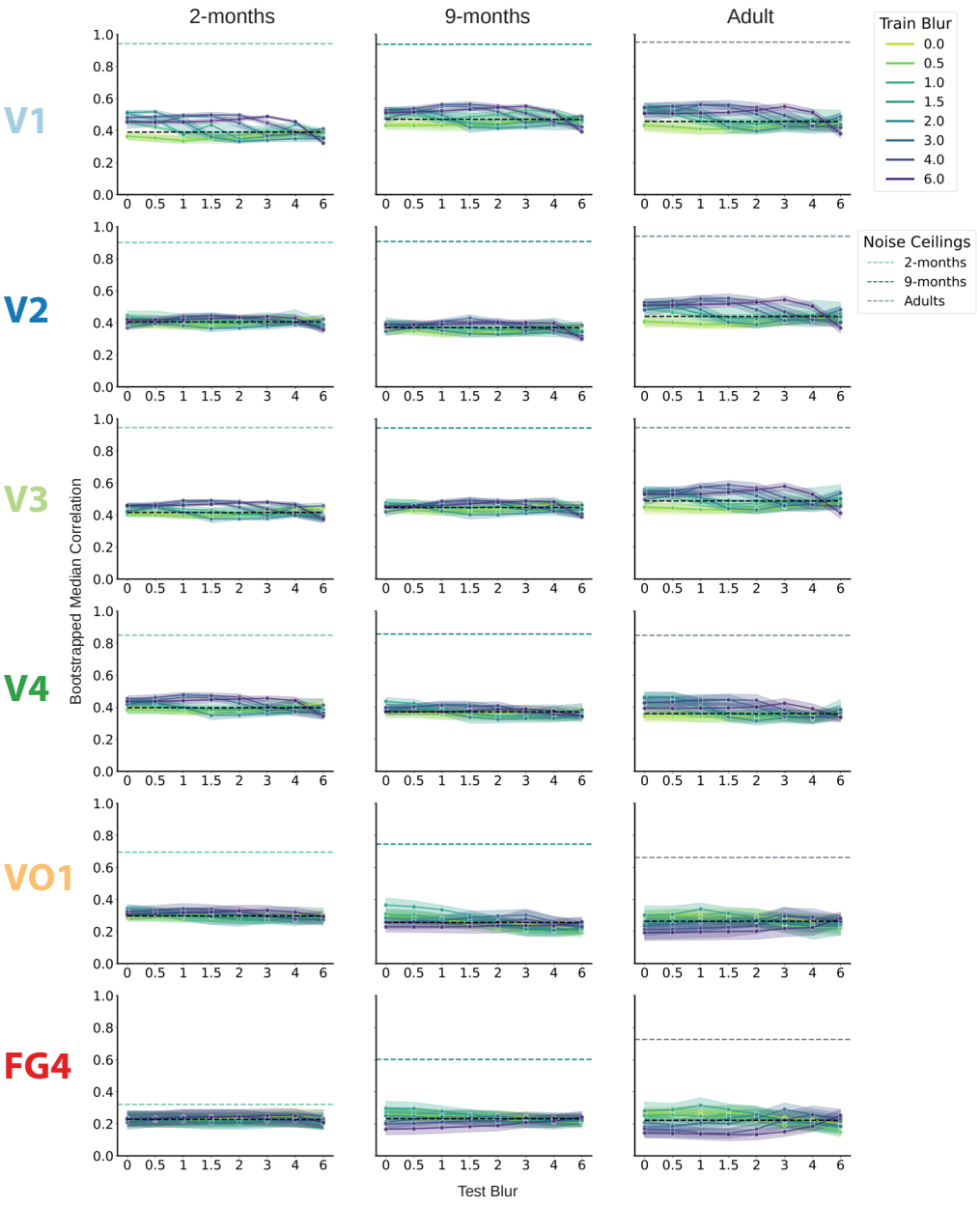

Figure A3.10 Reliability of object feature tuning across the developing ventral visual stream.

Alternative visualisation of the model–brain correspondences characterised by the multiscale tuning maps in (Chapter 3, Figure 3.2 and Figure A3.6). Each coloured line shows the bootstrapped median correlation value of the layer yielding the highest correlation with the brain (y-axis) for a given value of training blur (hue) across the range of testing blurs (x-axis). The error band depicts the 95% confidence interval of the correlation across bootstrap resamples. The thick black dashed line depicts the median

correlation of the maximum layer across all bootstraps for all combinations of train and test blur. The dashed green, teal and grey lines denote the split-half noise ceiling for each region for 2-month-olds, 9-month-olds and adults respectively, setting an upper bound on model correspondences. Correlation values were not scaled by the noise ceiling as this would artificially inflate estimates in later regions, FG4, where the split-half reliability was comparatively low.

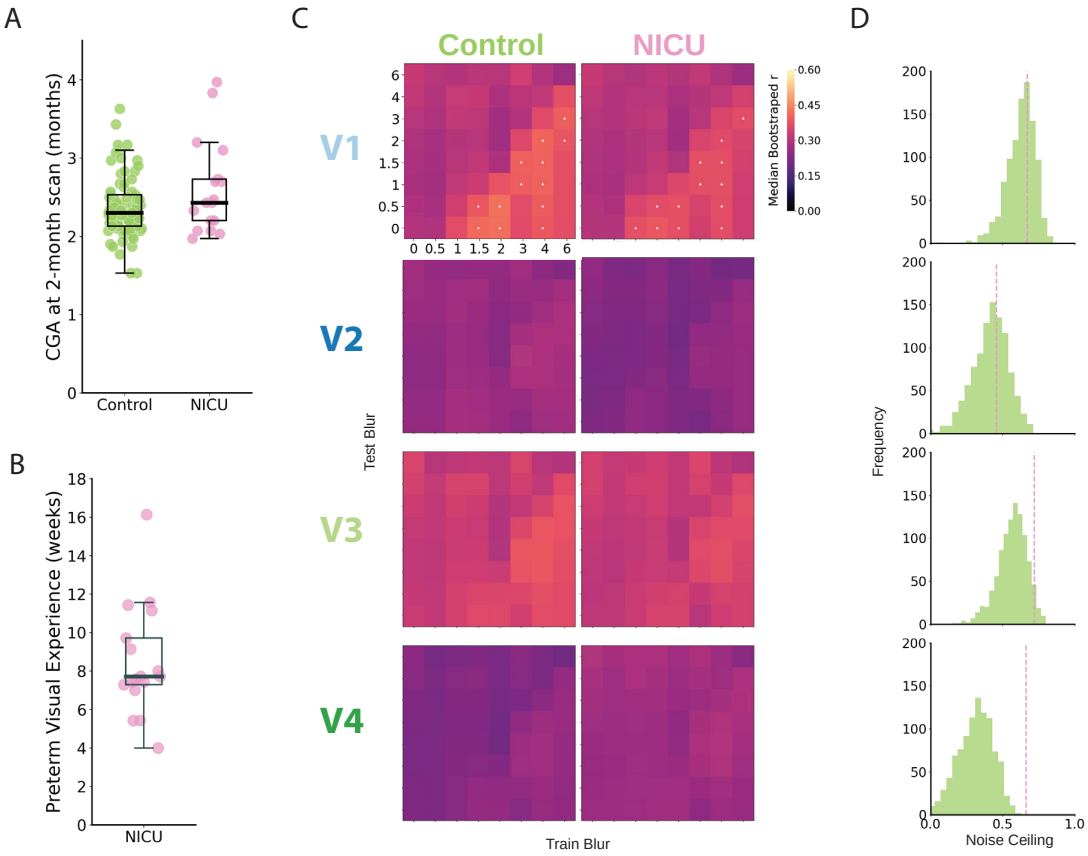

Figure A3.11 Prematurity was found not to affect the development of multiscale object feature tuning in the early visual cortex at 2 months.

To investigate whether the cumulative amount of early visual experience affected the development of object feature tuning in V1, preterm-born infants that spent time in the neonatal intensive care unit (NICU group) were compared to bootstrap samples of run-matched controls (control group). Both groups had a similar corrected gestational age at scan (A); however the NICU group were born an average of 8 weeks premature and therefore had an additional 8 weeks of postnatal visual experience at scan time (B). Multiscale object feature tuning between the groups was highly similar, indicating no effect of additional postnatal visual experience on the tuning maps at 2 months (C). The split-half noise ceiling for the NICU group (dashed pink line) was within the distribution of the noise ceilings for the bootstrapped samples of the run-matched controls (green distribution) for V1 and V2. Interestingly, in V3 and V4, the split-half noise ceiling was higher for the NICU group, indicating that there was less variability in object responses among this cohort in these regions.

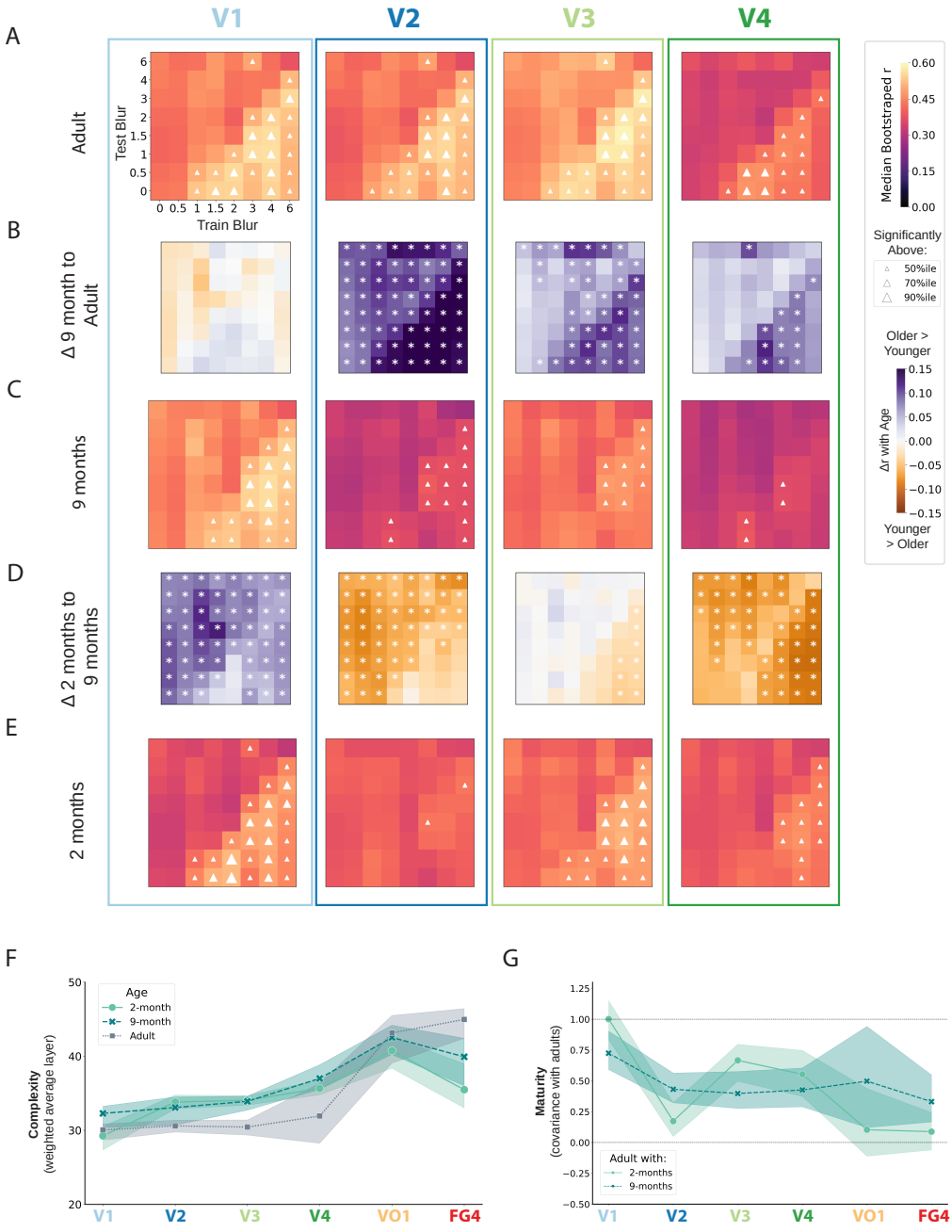

Figure A3.12 Replication of multiscale object feature tuning using a low-motion subset of infant runs.

The tuning patterns shown in Figure 3.2 were replicated when analyses were repeated using a low-motion subset of infant runs (see Chapter 3, Methods section 3.5.6.F). Tuning to object features across spatial scales is shown for adults (A), 9-month-olds (C), and 2-month-olds (E), with infant runs restricted to those with median FD below 0.5 mm. Difference maps (C,E), weighted average layer estimates (F), and maturity analyses

(G) were computed as in Figure 3.2. Overall, the developmental pattern was similar to that observed in the full dataset, with V1 appearing relatively mature by 2 months and later regions showing more protracted development. Analysis details and plotting conventions are consistent with those described for Figure 3.2.

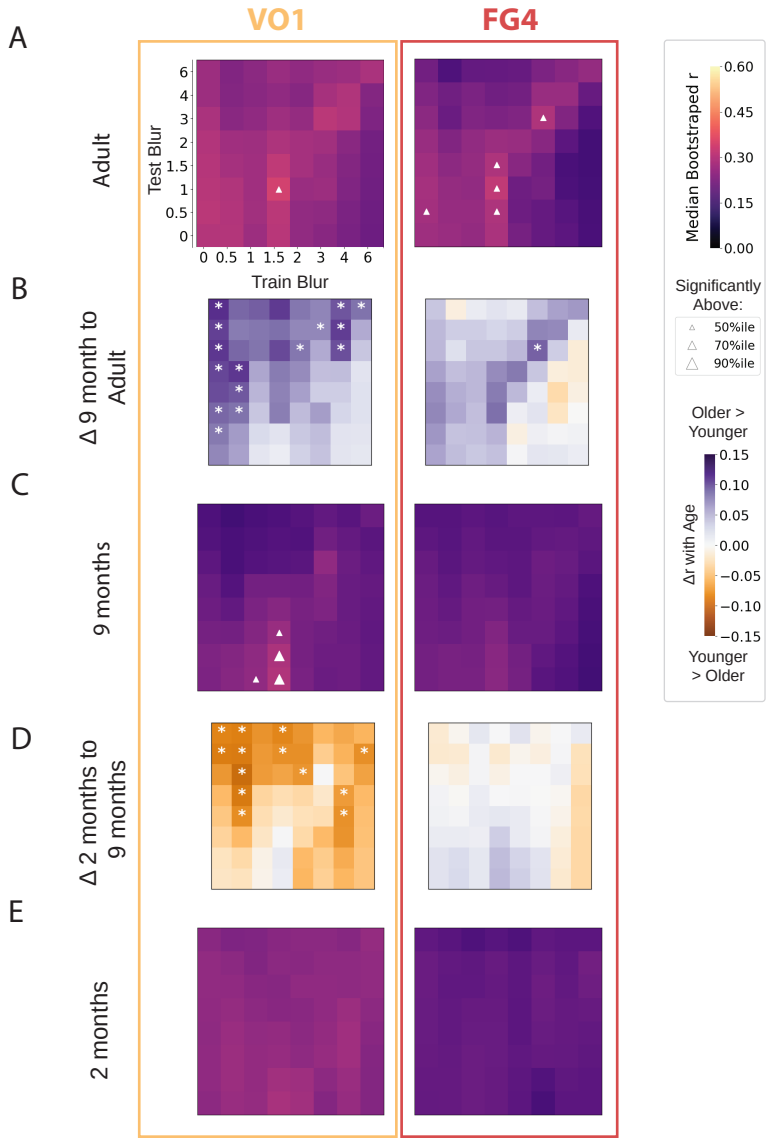

Figure A3.13 Replication of multiscale object feature tuning for later visual regions using a low motion subset of infant runs.

When the tuning patterns shown in Figure A3.6 were replicated using a low-motion subset of infant runs (see Chapter 3, Methods section 3.5.6.F) a similar pattern of results was observed.

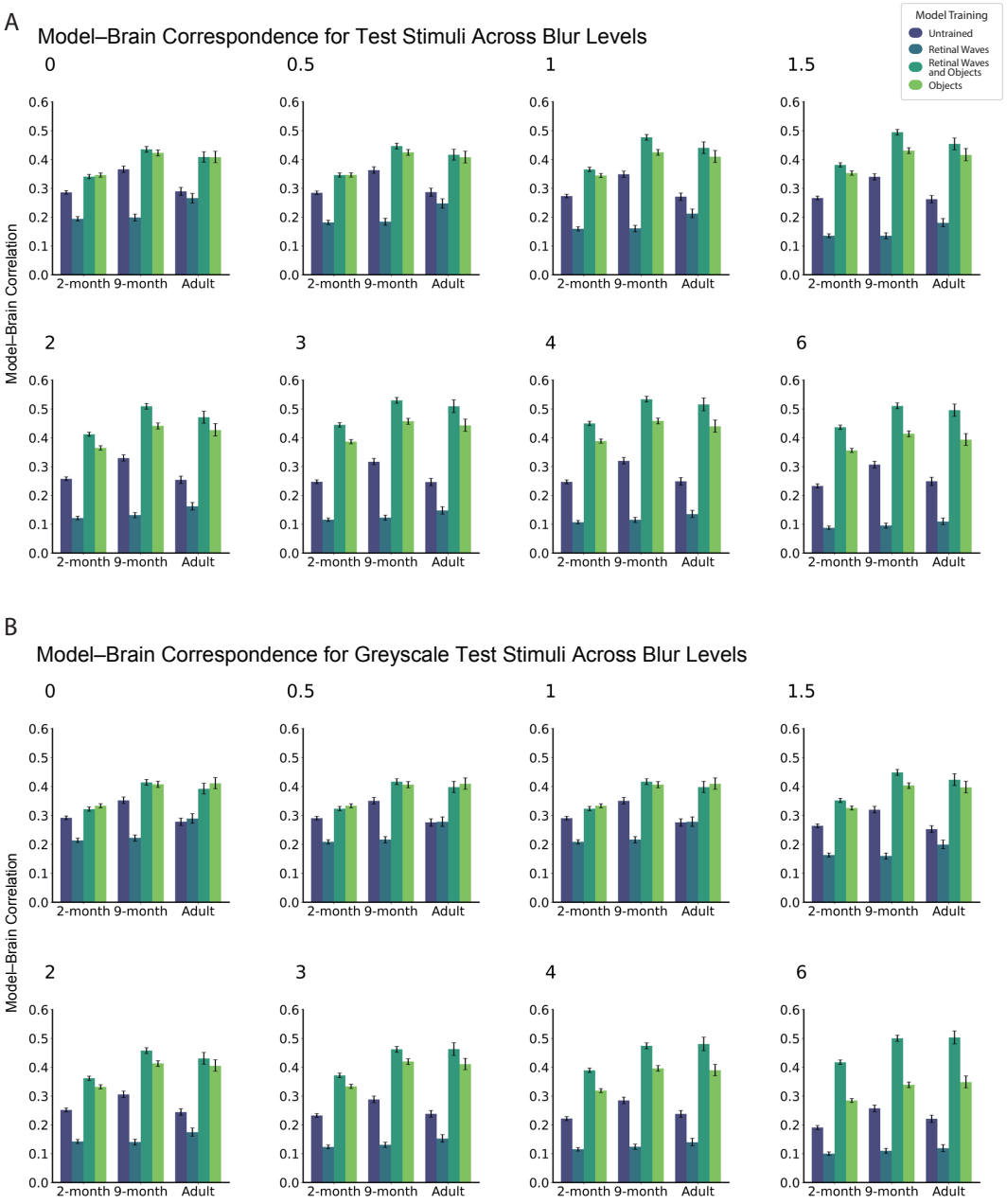

Figure A3.14 Correspondence between V1 and retinal-wave-trained models, tested on images across a range of blurs.

Across development training on retinal waves was found to be a poor model of V1 compared to training on objects. This suggests that, as young as 2 months, the features encoded by V1 correspond to those that are learned through object-specific training in DNNs. Each graph depicts the model-wise results when tested on the FOUNDCOG stimuli at the specified level of test blur. For sharp test images, pretraining on retinal

waves conferred no advantage over standard object training. At higher levels of test blur, however, pretraining on retinal waves resulted in increased representational correspondence with V1 (A). To determine whether colour might be driving the differences between retinal wave and object trained models, the analysis was repeated with greyscale versions of the test stimuli. Greyscale images produced a similar pattern of results (B), suggesting that colour is not driving the observed results.

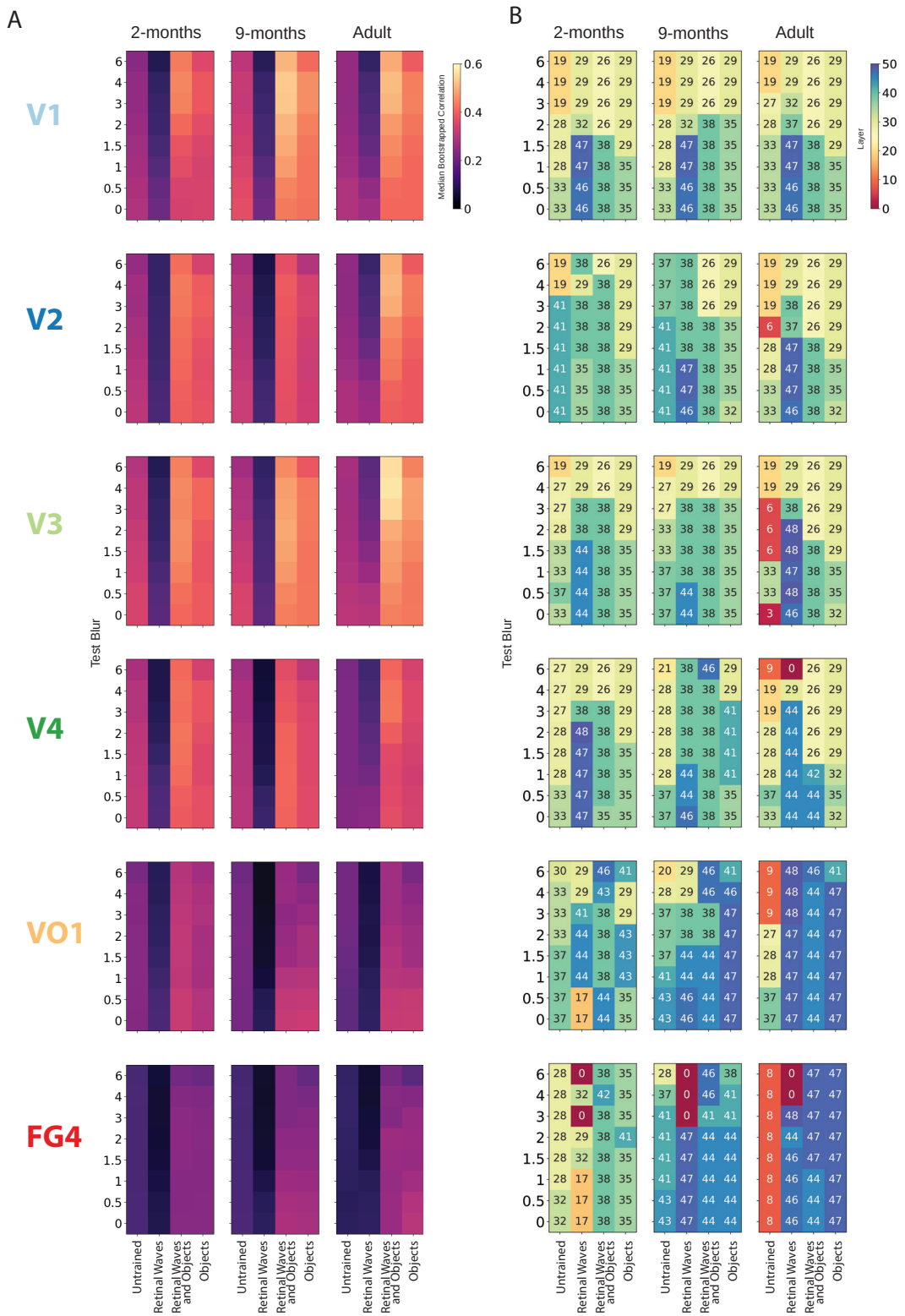

Figure A3.15 Model–brain correspondence for retinal-wave-trained models along the ventral visual stream.

Across the EVC, the pattern of model–brain correspondence was similar to that observed in V1 (Figure A3.14A) both for models pretrained on retinal waves and subsequently trained on objects and for models trained only on objects. In later object-selective regions, VO1 and FG4, however, retinal wave pretraining no longer conferred a representational advantage. The median bootstrapped correlation for the best-corresponding layer for each train–test combination is shown in (A). The bootstrap median layer associated with each train–test combination is shown in (B), analogous to the layers depicted in the multiscale tuning maps in Figure A3.4A.

Appendix 4

A4.1 Supplementary methods

A4.1.1 Empirically estimating equivalence between Gaussian blur and Butterworth cutoffs

We estimated Butterworth cutoff frequencies that produced frequency attenuation comparable to Gaussian blur with σ = 4 pixels (approximate low-pass boundary) and σ = 1 pixel (approximate high-pass boundary). Details of the Butterworth implementation are provided in the Methods (Section 4.2.1).

A4.1.2 Impulse image

A single-channel 224 × 224 image containing a single white central pixel (112, 112) on a black background was generated. This 'impulse' stimulus provides a simple way to characterise each filter's effect on spatial frequencies, as the filter's frequency attenuation can be read out directly from the amplitude spectrum of the filtered impulse image.

A4.1.3 Candidate filtering and spectral summary

The impulse image was filtered using Gaussian blur (σ = 4 or σ = 1) and a sweep of Butterworth filters spanning candidate cutoff frequencies. For each filtered image, we computed the Fourier transform and examined the Fourier magnitude spectrum. For comparison across filters, we extracted a 1D profile by taking a central horizontal slice of the 2D Fourier magnitude spectrum (values along $[x, 112]$), which summarises attenuation across spatial

frequencies along one axis. We denote this 1D slice as the Fourier magnitude spectrum $A(f)$, where f is spatial frequency along the horizontal axis (in cycles per pixel, cpp).

A4.1.3.A Estimating equivalent low-pass boundary (σ = 4)

To select the Butterworth cutoff corresponding to σ = 4, we compared the $A(f)$ from the Gaussian-filtered impulse image to the $A(f)$ produced by low-pass Butterworth filters across candidate cutoffs. The closest visual match in attenuation profile was obtained at 0.055 cpp.

A4.1.3.B Estimating equivalent high-pass boundary (σ = 1)

To select the Butterworth cutoff corresponding to the high-pass complement of a Gaussian blur with σ = 1, we constructed high-pass filtered images by subtracting the corresponding low-pass filtered impulse image from the original impulse image (high-pass = impulse − low-pass) for each candidate filter. We then computed Fourier magnitude spectra and extracted the same central horizontal slice, yielding $\boldsymbol{A(f)}$as above.

As Gaussian and Butterworth filters attenuate frequencies at different rates, we put the spectra on a common scale prior to comparison. Specifically, each 1D spectrum was normalised to its peak value ($max_f\{A(f)\}$) and converted to decibels relative to this peak (Equation A4.1).

$$A_{dB}(f) = 20log_{10}\left(\frac{A(f)}{max_f\{A(f)\}}\right) \quad \text{(A4.1)}$$

We then compared the $\sigma = 1$ Gaussian and Butterworth spectra at a common attenuation criterion of -10dB (relative to the peak), and selected the Butterworth cutoff whose

spectrum most closely matched the Gaussian at this threshold. This yielded a cutoff of 0.15 cpp.

A4.2 Supplementary figures

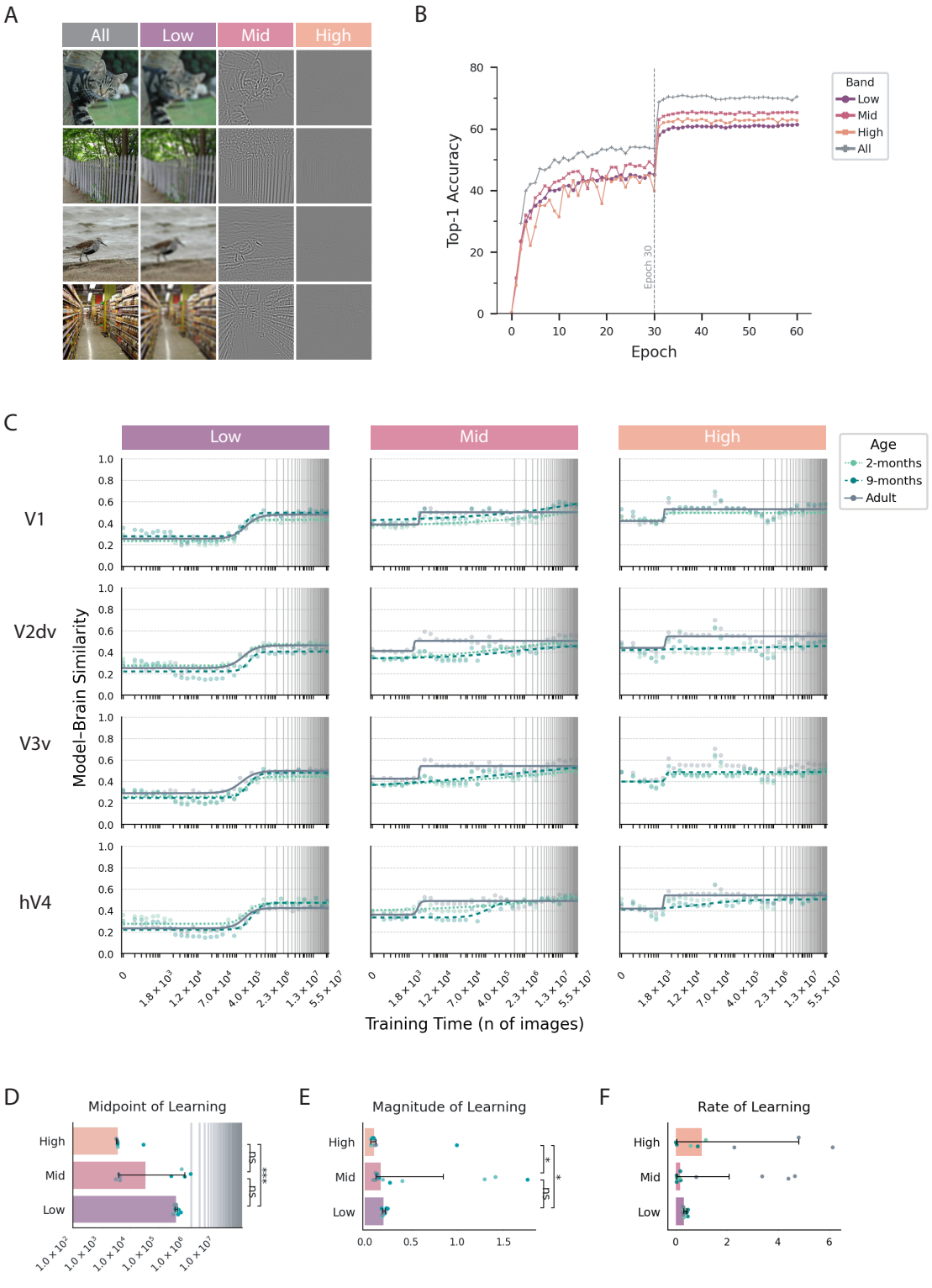

Figure A4.1 Preliminary characterisation of the band-wise trajectory of convergence with the early visual cortex across development to estimate approximate point of learning across bands.

Sample ImageNet stimuli decomposed into frequency bands containing low, mid and high spatial frequency information (A). Epoch-wise accuracy of models trained on bandpass filtered images for a duration of 60 epochs (B). Model–Brain correspondence was calculated between the developing EVC (V1–V4) and models trained on object features across spatial scales (low, mid or high) to estimate the approximate point of convergence with the EVC (C). To evaluate the dynamics of convergence, logistic curves were fit to each trajectory. This enables band-wise learning dynamics to be extrapolated, including the mid-point of learning (D), the magnitude of learning (E) and the rate of learning (F).

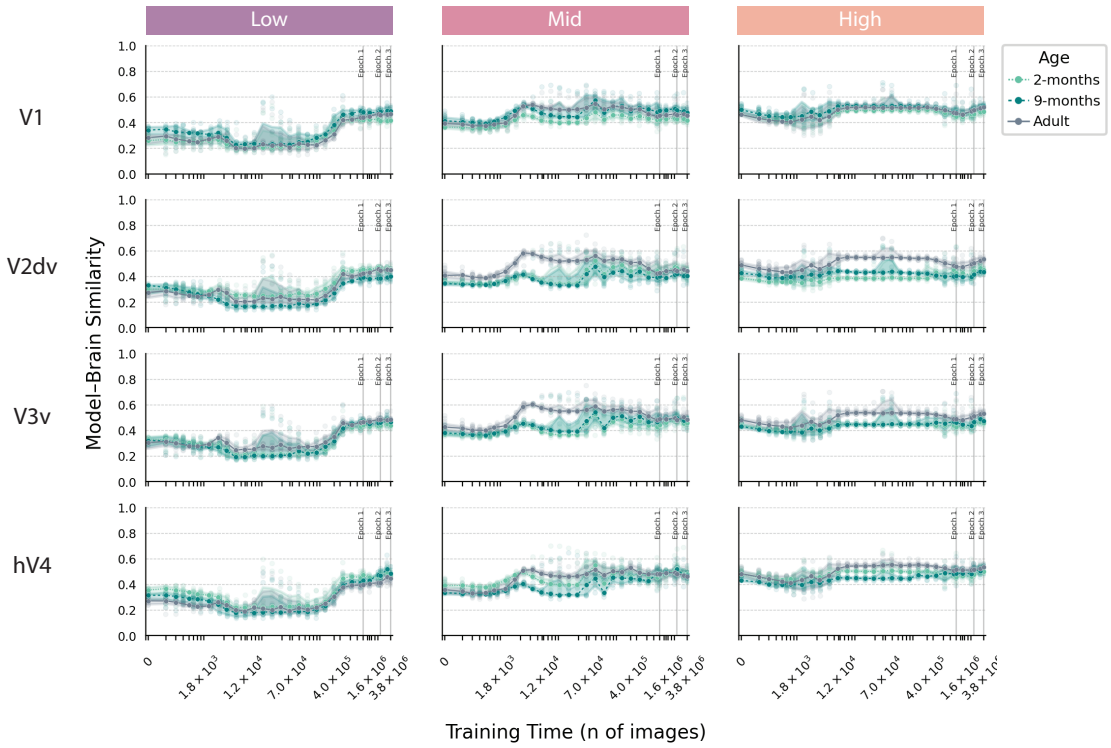

Figure A4.2 Band-wise trajectory of convergence with the early visual cortex across development.

Median and 95% confidence interval across models (n = 10 iterations per band).

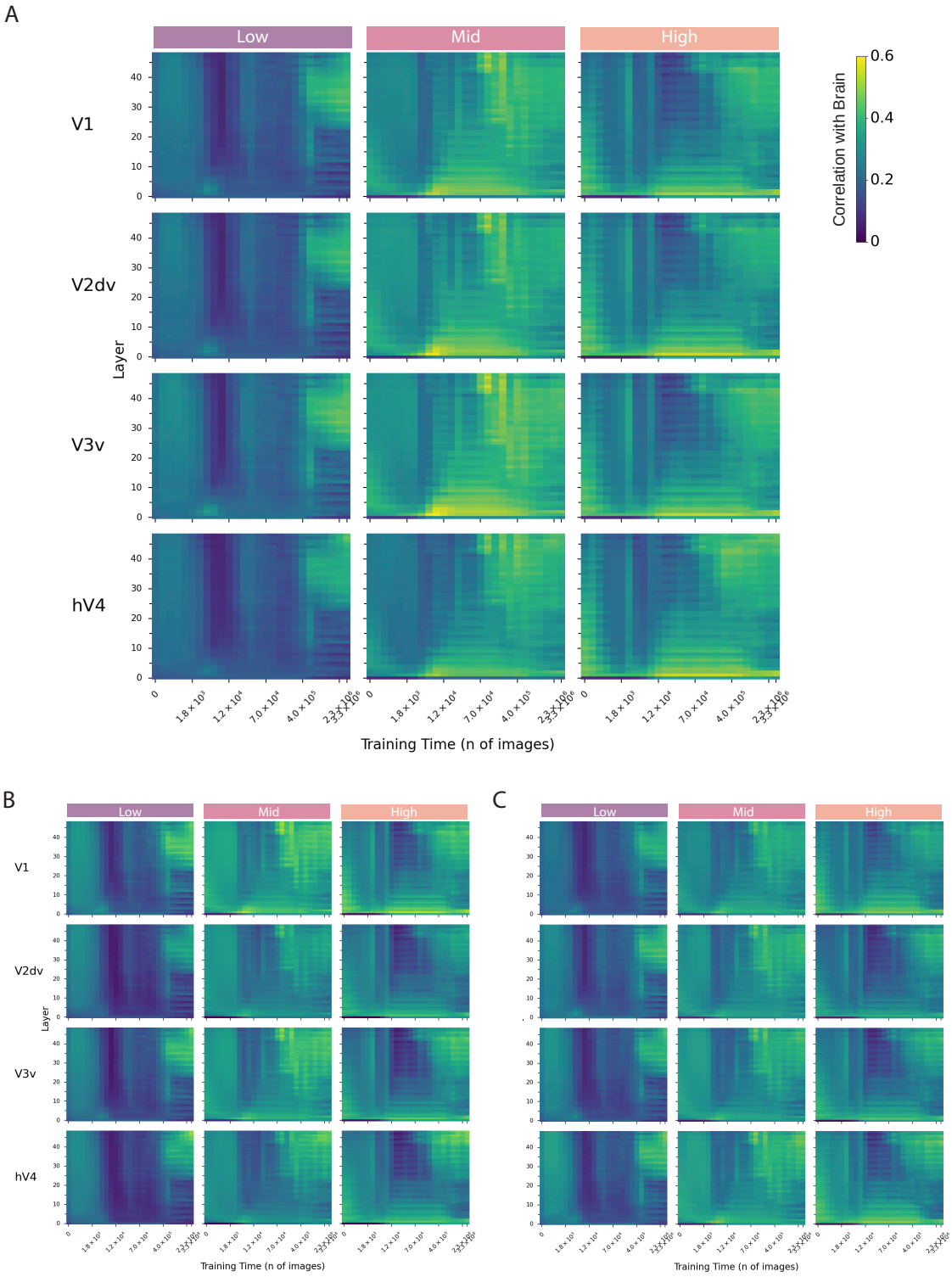

Figure A4.3 Layer-wise trajectories of convergence by band throughout the early visual cortex across development.

To visualise which layers were driving responses across the training period (3 epochs on a log scale), the layer-wise correspondence was visualised. The pattern of layer-wise

correspondence was found to be similar throughout the EVC across development; Adults (A), 9 months (B), 2 months (C)

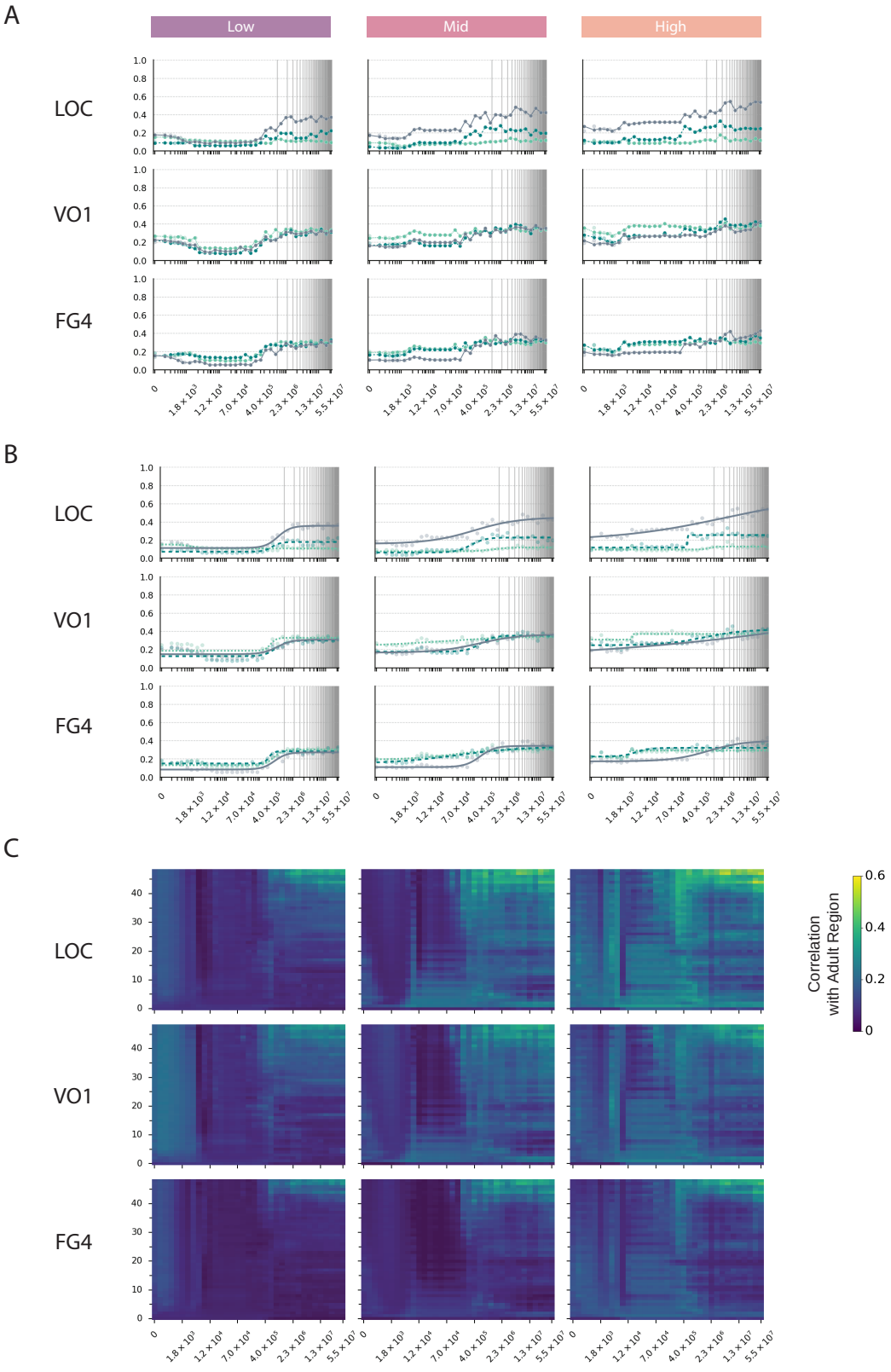

Figure A4.4 Preliminary characterisation of the band-wise trajectory of convergence with the later ventral visual stream regions across development.

Model–Brain correspondence was calculated between intermediate visual regions LOC and VO1, and a category selective region in the ventral-temporal cortex (FG4) and preliminary models trained for 60 epochs on object features across spatial scales (A). A logistic curve was fit to each trajectory (B). Layer-wise correspondence was visualised across model training revealing that correspondence was mediated by late model layers.

A4.3 Supplementary tables

Table A4.1 Kruskal–Wallis omnibus tests of spatial-frequency band effects on model–brain correspondence in untrained networks by ROI and age (BH–FDR corrected)

ROI	Age	df	H	p-value	Corrected p-value (corr)	Significant (corr)
V1	Adult	2	24.8155	0.0000	0.0000	Yes
V1	9 months	2	25.551	0.0000	0.0000	Yes
V1	2 months	2	25.551	0.0000	0.0000	Yes
V2dv	Adult	2	25.3006	0.0000	0.0000	Yes
V2dv	9 months	2	24.3123	0.0000	0.0000	Yes
V2dv	2 months	2	24.5806	0.0000	0.0000	Yes
V3v	Adult	2	21.0658	0.0000	0.0000	Yes
V3v	9 months	2	25.3006	0.0000	0.0000	Yes
V3v	2 months	2	21.5252	0.0000	0.0000	Yes
hV4	Adult	2	23.2981	0.0000	0.0000	Yes
hV4	9 months	2	19.7265	0.0001	0.0001	Yes
hV4	2 months	2	19.6206	0.0001	0.0001	Yes

Table A4.2 Post-hoc pairwise Mann–Whitney U tests comparing spatial-frequency bands on model–brain correspondence in untrained networks by ROI and age (Holm corrected)

ROI	Age	Comparison	U	p-value	p-value (corr)	Median difference	Significant (corr)
V1	Adult	low vs mid	0	0.0002	0.0005	0.0869	Yes
V1	Adult	low vs high	0	0.0002	0.0005	0.1371	Yes
V1	Adult	mid vs high	4	0.0006	0.0006	0.0502	Yes
V1	9 months	low vs mid	0	0.0002	0.0005	0.0529	Yes
V1	9 months	low vs high	0	0.0002	0.0005	0.1212	Yes
V1	9 months	mid vs high	1	0.0002	0.0005	0.0683	Yes
V1	2 months	low vs mid	0	0.0002	0.0005	0.0751	Yes
V1	2 months	low vs high	0	0.0002	0.0005	0.1529	Yes
V1	2 months	mid vs high	1	0.0002	0.0005	0.0778	Yes
V2dv	Adult	low vs mid	0	0.0002	0.0005	0.0814	Yes
V2dv	Adult	low vs high	0	0.0002	0.0005	0.1456	Yes
V2dv	Adult	mid vs high	2	0.0003	0.0005	0.0642	Yes
V2dv	9 months	low vs mid	5	0.0008	0.0008	0.0354	Yes
V2dv	9 months	low vs high	0	0.0002	0.0005	0.0945	Yes
V2dv	9 months	mid vs high	1	0.0002	0.0005	0.0591	Yes
V2dv	2 months	low vs mid	5	0.0008	0.0008	0.0326	Yes
V2dv	2 months	low vs high	0	0.0002	0.0005	0.0797	Yes
V2dv	2 months	mid vs high	0	0.0002	0.0005	0.0471	Yes
V3v	Adult	low vs mid	1	0.0002	0.0005	0.0688	Yes
V3v	Adult	low vs high	0	0.0002	0.0005	0.099	Yes
V3v	Adult	mid vs high	22	0.0376	0.0376	0.0301	Yes
V3v	9 months	low vs mid	0	0.0002	0.0005	0.0583	Yes
V3v	9 months	low vs high	0	0.0002	0.0005	0.102	Yes
V3v	9 months	mid vs high	2	0.0003	0.0005	0.0438	Yes
V3v	2 months	low vs mid	21	0.0312	0.0312	0.0116	Yes
V3v	2 months	low vs high	0	0.0002	0.0005	0.0838	Yes
V3v	2 months	mid vs high	0	0.0002	0.0005	0.0722	Yes
hV4	Adult	low vs mid	6	0.001	0.0012	0.0406	Yes
hV4	Adult	low vs high	0	0.0002	0.0005	0.0981	Yes
hV4	Adult	mid vs high	4	0.0006	0.0012	0.0575	Yes
hV4	9 months	low vs mid	38	0.3847	0.3847	0.0075	No
hV4	9 months	low vs high	0	0.0002	0.0005	0.0863	Yes
hV4	9 months	mid vs high	0	0.0002	0.0005	0.0787	Yes
hV4	2 months	low vs mid	31	0.162	0.162	0.0102	No
hV4	2 months	low vs high	0	0.0002	0.0005	0.0691	Yes
hV4	2 months	mid vs high	2	0.0003	0.0007	0.0588	Yes

Table A4.3 Channel-wise mean and standard deviation of bandpass-filtered ImageNet training images (RGB)

Band	ImageNet Mean [R,G,B]	ImageNet Standard Deviation [R,G,B]
All (Standard)	[0.485, 0.456, 0.406]	[0.229, 0.224, 0.225]
Low	[0.490, 0.467, 0.414]	[0.260, 0.253, 0.268]
Mid	[0.091, 0.089, 0.089]	[0.104, 0.102, 0.103]
High	[0.054, 0.054, 0.054]	[0.077, 0.077, 0.076]

Table A4.4 Preliminary band-wise logistic fit parameters across 60 training epochs by ROI and age (1 iteration per band)

ROI	Train band	Age	N	Lower asymptote (y_{min})	Upper asymptote (y_{max})	Inflection (x_0)	Slope (k)	Logistic R^2 (adj.)	Log-linear R^2 (adj.)	ΔR^2 (adj.)	Best model	Best R^2 (adj.)
V1	Low	Adult	35	0.2543	0.4799	3.1808	5.6131	0.9601	0.7526	0.2075	logistic	0.9601
V1	Low	9 months	35	0.2783	0.4971	3.1543	8.786	0.8758	0.5555	0.3203	logistic	0.8758
V1	Low	2 months	35	0.2335	0.4315	3.082	8.5146	0.9323	0.6506	0.2817	logistic	0.9323
V2dv	Low	Adult	35	0.2535	0.4645	3.1204	5.4814	0.9481	0.765	0.1831	logistic	0.9481
V2dv	Low	9 months	35	0.2218	0.4065	3.2726	9.0176	0.7119	0.3349	0.377	logistic	0.7119
V2dv	Low	2 months	35	0.2767	0.4744	3.1779	6.6405	0.9388	0.6677	0.2711	logistic	0.9388
V3v	Low	Adult	35	0.2903	0.4976	3.077	5.3229	0.9533	0.7897	0.1636	logistic	0.9533
V3v	Low	9 months	35	0.2462	0.4804	3.2466	7.4471	0.889	0.5579	0.331	logistic	0.889
V3v	Low	2 months	35	0.2571	0.4437	3.2149	6.5879	0.8178	0.4986	0.3191	logistic	0.8178
hV4	Low	Adult	35	0.2346	0.4234	3.1555	6.6028	0.942	0.6708	0.2712	logistic	0.942
hV4	Low	9 months	35	0.2217	0.4717	3.3474	7.5829	0.8127	0.4436	0.3691	logistic	0.8127
hV4	Low	2 months	35	0.2769	0.4735	3.2049	6.4402	0.7753	0.4394	0.3359	logistic	0.7753
V1	Mid	Adult	35	0.387	0.5017	1.2265	162.3846	0.5908	0.3755	0.2152	logistic	0.5908
V1	Mid	9 months	35	0.4113	2.1714	10.3301	0.4462	0.5292	0.5339	-0.0047	loglinear	0.5339
V1	Mid	2 months	35	0.3595	1.656	10.3301	0.4119	0.5602	0.5688	-0.0085	loglinear	0.5688
V2dv	Mid	Adult	35	0.413	0.5062	1.0856	145.0195	0.3685	0.1002	0.2683	logistic	0.3685
V2dv	Mid	9 months	35	0.3436	0.476	3.6861	1.1427	0.5134	0.504	0.0094	logistic	0.5134
V2dv	Mid	2 months	35	0.3329	0.5297	3.3721	0.9011	0.7072	0.7063	0.001	logistic	0.7072
V3v	Mid	Adult	35	0.4249	0.5428	1.2228	149.4464	0.576	0.2697	0.3062	logistic	0.576
V3v	Mid	9 months	35	0.3291	0.6031	3.4799	0.5255	0.6841	0.6998	-0.0157	loglinear	0.6998
V3v	Mid	2 months	35	0.3328	0.7403	6.3535	0.3732	0.5313	0.5481	-0.0168	loglinear	0.5481
hV4	Mid	Adult	35	0.3616	0.4885	1.1851	25.2865	0.6509	0.4839	0.167	logistic	0.6509

hV4	Mid	9 months	35	0.3353	0.4901	3.0122	4.6865	0.815	0.6934	0.1216	logistic	0.815
hV4	Mid	2 months	35	0.3854	1.799	10.3301	0.4247	0.5148	0.5081	0.0067	logistic	0.5148
V1	High	Adult	35	0.4202	0.5297	1.1037	224.1783	0.4975	0.0807	0.4168	logistic	0.4975
V1	High	2 months	35	0.4285	0.4957	1.1542	70.0526	0.3454	0.0405	0.3049	logistic	0.3454
V2dv	High	Adult	35	0.4409	0.5483	1.1444	85.2808	0.4005	0.034	0.3666	logistic	0.4005
V2dv	High	9 months	35	-0.0023	0.9939	10.33	0.0299	-0.0934	-0.0251	-0.0683	loglinear	-0.0251
V3v	High	9 months	35	0.3986	0.4865	1.1468	39.1809	0.2526	0.0711	0.1815	logistic	0.2526
V3v	High	2 months	35	0.3997	0.4684	1.1552	33.9385	0.1529	0.0363	0.1166	logistic	0.1529
hV4	High	Adult	35	0.4177	0.5412	1.0971	155.784	0.6323	0.1422	0.4902	logistic	0.6323
hV4	High	9 months	35	0.4058	0.5046	2.0465	1.6866	0.3794	0.3932	-0.0138	loglinear	0.3932

Table A4.5 Kruskal–Wallis omnibus tests of band effects on preliminary learning dynamics across 60 training epochs by ROI and age (1 iteration per band; BH–FDR corrected)

Metric	df	H	p-value	n high	n mid	n low	n total
Midpoint of Learning	2	25.6551	0	11	12	12	35
Magnitude of Learning	2	21.2207	0	12	12	12	36
Rate of Learning	2	4.1186	0.1275	12	12	12	36

Table A4.6 Post-hoc pairwise Mann–Whitney U tests comparing spatial-frequency bands for preliminary learning dynamics across 60 training epochs by ROI and age (1 iteration per band; Holm corrected)

Metric	Band 1	Band 2	p-value	Corrected p-value(Holm)	Significant (corr)
Midpoint of Learning	High	Mid	0.0401	0.0802	No
Midpoint of Learning	High	Low	0	0.0001	Yes
Midpoint of Learning	Mid	Low	0.3837	0.3837	No
Magnitude of Learning	High	Mid	0.0096	0.0191	Yes
Magnitude of Learning	High	Low	0.0041	0.0123	Yes
Magnitude of Learning	Mid	Low	0.665	0.665	No

Table A4.7 Band-wise logistic fit parameters early in training (epochs 0–3) by ROI and age (10 model iterations per band)

ROI	Train band	Age	N	Lower asymptote (y_{min})	Upper asymptote (y_{max})	Inflection (x_0)	Slope (k)	Logistic R^2 (adj.)	Log-linear R^2 (adj.)	ΔR^2 (adj.)	Best model	Best R^2 (adj.)
V1	Low	Adult	270	0.2529	0.4451	3.2238	16.5667	0.6445	0.3053	0.3392	logistic	0.6445
V1	Low	9 months	270	0.2921	0.4798	3.1881	20.8732	0.5918	0.2219	0.3699	logistic	0.5918
V1	Low	2 months	270	0.246	0.4204	3.1423	13.016	0.6643	0.3602	0.3041	logistic	0.6643
V2dv	Low	Adult	270	0.2525	0.4288	3.224	20.3557	0.5904	0.2781	0.3123	logistic	0.5904
V2dv	Low	9 months	270	0.229	0.3843	3.2402	19.4266	0.4888	0.084	0.4048	logistic	0.4888
V2dv	Low	2 months	270	0.2825	0.4517	3.2041	21.4092	0.6673	0.3343	0.333	logistic	0.6673
V3v	Low	Adult	270	0.2907	0.4658	3.2218	15.3984	0.5921	0.3254	0.2667	logistic	0.5921
V3v	Low	9 months	270	0.2591	0.4593	3.2283	16.5978	0.6239	0.2361	0.3878	logistic	0.6239
V3v	Low	2 months	270	0.2688	0.4327	3.2223	27.7413	0.5361	0.1665	0.3696	logistic	0.5361
hV4	Low	Adult	270	0.2371	0.4126	3.2351	13.8251	0.6305	0.305	0.3254	logistic	0.6305
hV4	Low	9 months	270	0.2354	0.4547	3.2785	11.3556	0.6415	0.1912	0.4503	logistic	0.6415
hV4	Low	2 months	270	0.2858	0.4626	3.2242	26.2364	0.5418	0.1495	0.3923	logistic	0.5418
V1	Mid	Adult	270	0.386	0.5114	1.2303	151.2659	0.6027	0.2212	0.3815	logistic	0.6027
V1	Mid	9 months	270	0.4102	0.5094	1.1306	18.1604	0.3559	0.2133	0.1426	logistic	0.3559
V1	Mid	2 months	270	0.3624	0.4407	1.1662	17.7222	0.41	0.238	0.1721	logistic	0.41
V2dv	Mid	Adult	270	0.4036	0.5206	1.1302	20.9383	0.4649	0.0731	0.3918	logistic	0.4649
V2dv	Mid	9 months	270	0.339	0.4212	1.3577	2.2103	0.1484	0.1267	0.0217	logistic	0.1484
V2dv	Mid	2 months	270	0.3461	0.4643	1.7654	2.1628	0.3035	0.2822	0.0213	logistic	0.3035
V3v	Mid	Adult	270	0.4188	0.5555	1.2288	170.4561	0.6	0.1871	0.4129	logistic	0.6
V3v	Mid	9 months	270	0.3674	0.4933	1.501	2.2869	0.3314	0.2944	0.037	logistic	0.3314
V3v	Mid	2 months	270	0.353	0.463	1.3053	1.8636	0.2229	0.206	0.0169	logistic	0.2229
hV4	Mid	Adult	270	0.3484	0.4919	1.2046	18.6135	0.7356	0.4312	0.3044	logistic	0.7356

hV4	Mid	9 months	270	0.3325	0.515	2.806	1.3004	0.401	0.3903	0.0107	logistic	0.401
hV4	Mid	2 months	270	0.3892	0.4748	1.0865	22.6073	0.2373	0.186	0.0514	logistic	0.2373
V1	High	Adult	270	0.4325	0.5128	1.5187	69.8351	0.3911	0.1901	0.201	logistic	0.3911
V1	High	9 months	270	0.4696	0.5231	1.5098	135.427	0.201	0.0561	0.1449	logistic	0.201
V1	High	2 months	270	0.4345	0.4868	1.5206	134.3463	0.3195	0.1043	0.2153	logistic	0.3195
V2dv	High	Adult	270	0.4619	0.5347	1.5139	144.6092	0.3374	0.1096	0.2278	logistic	0.3374
V2dv	High	9 months	270	0.4099	0.4312	1.2452	10.5705	0.0261	-0.0068	0.0329	logistic	0.0261
V2dv	High	2 months	270	0.3673	0.7375	6.0716	0.6574	0.1423	0.1347	0.0076	logistic	0.1423
V3v	High	Adult	270	0.4526	0.5319	1.5161	128.7167	0.3501	0.1594	0.1907	logistic	0.3501
V3v	High	9 months	270	0.4117	0.4633	1.5895	6.6057	0.175	0.134	0.0411	logistic	0.175
V3v	High	2 months	270	0.4183	0.45	1.4994	118.2947	0.1143	0.0482	0.066	logistic	0.1143
hV4	High	Adult	270	0.4462	0.5384	1.5485	21.1467	0.4774	0.2229	0.2545	logistic	0.4774
hV4	High	9 months	270	0.4132	0.4987	2.5692	1.4359	0.3456	0.3232	0.0224	logistic	0.3456
hV4	High	2 months	270	0.4559	0.4976	1.4827	50.3881	0.1971	0.0565	0.1406	logistic	0.1971

Table A4.8 Kruskal–Wallis omnibus tests of band effects on early learning dynamics (epochs 0–3) by ROI and age (10 model iterations per band; BH–FDR corrected)

Metric	df	H	p-value	n high	n mid	n low	n total	Significant
Midpoint of Learning	2	25.6551	0	11	12	12	35	Yes
Magnitude of Learning	2	21.2207	0	12	12	12	36	Yes
Rate of Learning	2	4.1186	0.1275	12	12	12	36	No

Table A4.9 Post-hoc pairwise Mann–Whitney U tests comparing spatial-frequency bands for early learning dynamics (epochs 0–3) by ROI and age (10 model iterations per band; Holm corrected)

Metric	Band 1	Band 2	p-value	Corrected p-value (Holm)	Significant (corr)
Midpoint of Learning	High	Mid	0.015	0.015	Yes
Midpoint of Learning	High	Low	0.0001	0.0001	Yes
Midpoint of Learning	Mid	Low	0	0.0001	Yes
Magnitude of Learning	High	Mid	0.0024	0.0024	Yes
Magnitude of Learning	High	Low	0.0006	0.0012	Yes
Magnitude of Learning	Mid	Low	0.0002	0.0007	Yes

Table A4.10 Within-network paired Wilcoxon signed-rank tests of change in model–brain correspondence from an untrained starting point to the endpoint of the critical training window (epoch 0 vs 3) by ROI, age, and band (10 model iterations per band; BH–FDR corrected across 36 tests)

ROI	Age	Band	N pairs	W+	p-value (corr)	Significant (corr)	Rank-biserial r	Median difference	Mean difference
V1	Adult	High	10	0	0.002	Yes	-1	0.1073	0.1117
V1	Adult	Low	10	0	0.002	Yes	-1	0.1848	0.1826
V1	Adult	Mid	10	0	0.002	Yes	-1	0.0955	0.0983
V1	9 months	High	10	0	0.002	Yes	-1	0.0596	0.065
V1	9 months	Low	10	0	0.002	Yes	-1	0.1321	0.1338
V1	9 months	Mid	10	0	0.002	Yes	-1	0.0824	0.0812
V1	2 months	High	10	0	0.002	Yes	-1	0.0574	0.0569
V1	2 months	Low	10	0	0.002	Yes	-1	0.1295	0.1288
V1	2 months	Mid	10	0	0.002	Yes	-1	0.065	0.0631
V2dv	Adult	High	10	0	0.002	Yes	-1	0.1107	0.1146
V2dv	Adult	Low	10	0	0.002	Yes	-1	0.1691	0.1724
V2dv	Adult	Mid	10	0	0.002	Yes	-1	0.0961	0.0899
V2dv	9 months	High	10	0	0.002	Yes	-1	0.0279	0.0359
V2dv	9 months	Low	10	0	0.002	Yes	-1	0.0755	0.0758
V2dv	9 months	Mid	10	2	0.0059	Yes	-0.9273	0.0526	0.044
V2dv	2 months	High	10	0	0.002	Yes	-1	0.0509	0.0516
V2dv	2 months	Low	10	0	0.002	Yes	-1	0.1352	0.131
V2dv	2 months	Mid	10	0	0.002	Yes	-1	0.0966	0.0816
V3v	Adult	High	10	0	0.002	Yes	-1	0.12	0.124
V3v	Adult	Low	10	0	0.002	Yes	-1	0.1715	0.1685
V3v	Adult	Mid	10	0	0.002	Yes	-1	0.1155	0.1081
V3v	9 months	High	10	0	0.002	Yes	-1	0.059	0.0612
V3v	9 months	Low	10	0	0.002	Yes	-1	0.144	0.1423
V3v	9 months	Mid	10	0	0.002	Yes	-1	0.1144	0.0972
V3v	2 months	High	10	0	0.002	Yes	-1	0.0397	0.0397
V3v	2 months	Low	10	0	0.002	Yes	-1	0.1002	0.1018
V3v	2 months	Mid	10	0	0.002	Yes	-1	0.1007	0.0927
hV4	Adult	High	10	0	0.002	Yes	-1	0.1378	0.1371
hV4	Adult	Low	10	0	0.002	Yes	-1	0.1565	0.1515
hV4	Adult	Mid	10	0	0.002	Yes	-1	0.1362	0.1344
hV4	9 months	High	10	0	0.002	Yes	-1	0.1068	0.0996
hV4	9 months	Low	10	0	0.002	Yes	-1	0.1438	0.1472
hV4	9 months	Mid	10	0	0.002	Yes	-1	0.1337	0.1257
hV4	2 months	High	10	0	0.002	Yes	-1	0.0611	0.0604
hV4	2 months	Low	10	0	0.002	Yes	-1	0.1027	0.1015
hV4	2 months	Mid	10	0	0.002	Yes	-1	0.1011	0.0943

Table A4.11 Kruskal–Wallis omnibus tests of spatial-frequency band effects on model–brain correspondence at the endpoint of the critical training window (epoch 3) by ROI and age (BH–FDR corrected)

ROI	Age	df	H	p-value	Corrected p-value (corr)	Significant (corr)
V1	Adult	2	18.0335	0.0001	0.0004	Yes
V1	9 months	2	12.1884	0.0023	0.0045	Yes
V1	2 months	2	18.2116	0.0001	0.0004	Yes
V2dv	Adult	2	19.8606	0	0.0003	Yes
V2dv	9 months	2	16.7458	0.0002	0.0006	Yes
V2dv	2 months	2	2.2529	0.3242	0.3242	No
V3v	Adult	2	11.6206	0.003	0.0051	Yes
V3v	9 months	2	2.48	0.2894	0.3157	No
V3v	2 months	2	6.6916	0.0352	0.0423	Yes
hV4	Adult	2	20.6116	0	0.0003	Yes
hV4	9 months	2	8.5084	0.0142	0.0189	Yes
hV4	2 months	2	9.2929	0.0096	0.0144	Yes

Table A4.12 Post-hoc pairwise Mann–Whitney U tests comparing spatial-frequency bands on model–brain correspondence at the endpoint of the critical training window (epoch 3) by ROI and age (Holm corrected)

ROI	Age	Comparison	U	p-value	p-value (corr)	Median difference	Significant (corr)
V1	Adult	low vs mid	68	0.1859	0.1859	-0.0124	No
V1	Adult	low vs high	4	0.0006	0.0012	0.0645	Yes
V1	Adult	mid vs high	2	0.0003	0.001	0.0769	Yes
V1	9 months	low vs mid	58	0.5708	0.5708	-0.0109	No
V1	9 months	low vs high	9	0.0022	0.0066	0.0485	Yes
V1	9 months	mid vs high	12	0.0046	0.0092	0.0594	Yes
V1	2 months	low vs mid	51	0.9698	0.9698	0.0017	No
V1	2 months	low vs high	0	0.0002	0.0005	0.0808	Yes
V1	2 months	mid vs high	3	0.0004	0.0009	0.0791	Yes
V2dv	Adult	low vs mid	64	0.3075	0.3075	-0.0114	No
V2dv	Adult	low vs high	0	0.0002	0.0005	0.0988	Yes
V2dv	Adult	mid vs high	0	0.0002	0.0005	0.1102	Yes
V2dv	9 months	low vs mid	49	0.9698	0.9698	0.0033	No
V2dv	9 months	low vs high	3	0.0004	0.0013	0.0373	Yes
V2dv	9 months	mid vs high	4	0.0006	0.0013	0.0341	Yes
V3v	Adult	low vs mid	45	0.7337	0.7337	0.0011	No
V3v	Adult	low vs high	9	0.0022	0.0066	0.0575	Yes
V3v	Adult	mid vs high	14	0.0073	0.0146	0.0564	Yes
V3v	2 months	low vs mid	28	0.1041	0.2082	0.0206	No
V3v	2 months	low vs high	15	0.0091	0.0273	0.0226	Yes
V3v	2 months	mid vs high	44	0.6776	0.6776	0.0019	No
hV4	Adult	low vs mid	21	0.0312	0.0312	0.0179	Yes
hV4	Adult	low vs high	0	0.0002	0.0005	0.0944	Yes
hV4	Adult	mid vs high	3	0.0004	0.0009	0.0764	Yes
hV4	9 months	low vs mid	62	0.3847	0.3847	-0.0155	No
hV4	9 months	low vs high	21	0.0312	0.0624	0.0503	No
hV4	9 months	mid vs high	15	0.0091	0.0273	0.0658	Yes
hV4	2 months	low vs mid	35	0.273	0.273	0.0115	No
hV4	2 months	low vs high	10	0.0028	0.0085	0.0309	Yes
hV4	2 months	mid vs high	26	0.0757	0.1513	0.0193	No

Table A4.13 Band-specific differences between infants and adults in model–brain correspondence by ROI (exact sign-flip permutation test; BH–FDR corrected)

ROI	Band	Age	N	Mean effect	95% CI	p-value (corr)	Significant (corr)
V1	High	Adult – 9 months	10	0.0101	[0.0047, 0.0152]	0.0445	Yes
V1	Low	Adult – 9 months	10	-0.0037	[-0.0118, 0.0033]	0.3932	No
V1	Mid	Adult – 9 months	10	-0.0064	[-0.0156, 0.0026]	0.2501	No
V1	High	Adult – 2 months	10	-0.0057	[-0.0105, -0.0014]	0.06	No
V1	Low	Adult – 2 months	10	0.01	[0.0032, 0.0169]	0.041	Yes
V1	Mid	Adult – 2 months	10	-0.0043	[-0.0097, 0.0029]	0.3009	No
V2dv	High	Adult – 9 months	10	0.0327	[0.0276, 0.0385]	0.0117	Yes
V2dv	Low	Adult – 9 months	10	-0.0107	[-0.0197, -0.0003]	0.1127	No
V2dv	Mid	Adult – 9 months	10	-0.022	[-0.0336, -0.0116]	0.0205	Yes
V2dv	High	Adult – 2 months	10	0.0636	[0.0505, 0.0791]	0.0117	Yes
V2dv	Low	Adult – 2 months	10	-0.0388	[-0.0489, -0.0275]	0.0117	Yes
V2dv	Mid	Adult – 2 months	10	-0.0248	[-0.0431, -0.0077]	0.0585	No
V3v	High	Adult – 9 months	10	0.0322	[0.0231, 0.0428]	0.0117	Yes
V3v	Low	Adult – 9 months	10	-0.0156	[-0.0281, 0.0018]	0.1158	No
V3v	Mid	Adult – 9 months	10	-0.0166	[-0.0338, 0.0002]	0.1346	No
V3v	High	Adult – 2 months	10	0.0275	[0.0182, 0.0369]	0.0117	Yes
V3v	Low	Adult – 2 months	10	-0.0092	[-0.0220, 0.0084]	0.3353	No
V3v	Mid	Adult – 2 months	10	-0.0184	[-0.0405, 0.0012]	0.1886	No
hV4	High	Adult – 9 months	10	0.022	[0.0066, 0.0360]	0.0566	No
hV4	Low	Adult – 9 months	10	-0.0368	[-0.0468, -0.0275]	0.0117	Yes
hV4	Mid	Adult – 9 months	10	0.0148	[0.0014, 0.0296]	0.1127	No
hV4	High	Adult – 2 months	10	0.0377	[0.0256, 0.0485]	0.0117	Yes
hV4	Low	Adult – 2 months	10	-0.0317	[-0.0439, -0.0200]	0.0117	Yes
hV4	Mid	Adult – 2 months	10	-0.006	[-0.0238, 0.0112]	0.5395	No