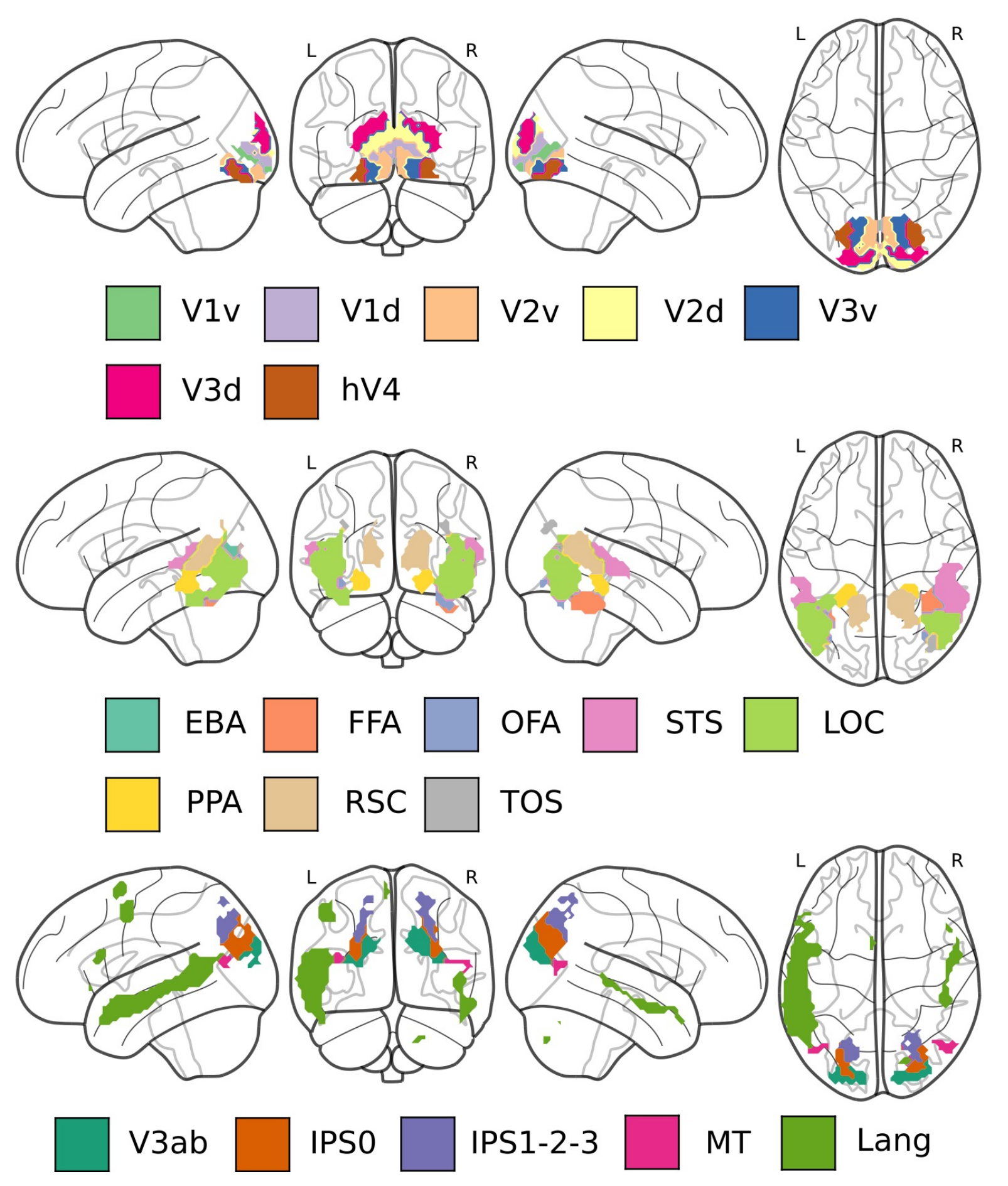
Investigating the role of modality and training objective on representational alignment between transformers and the brain

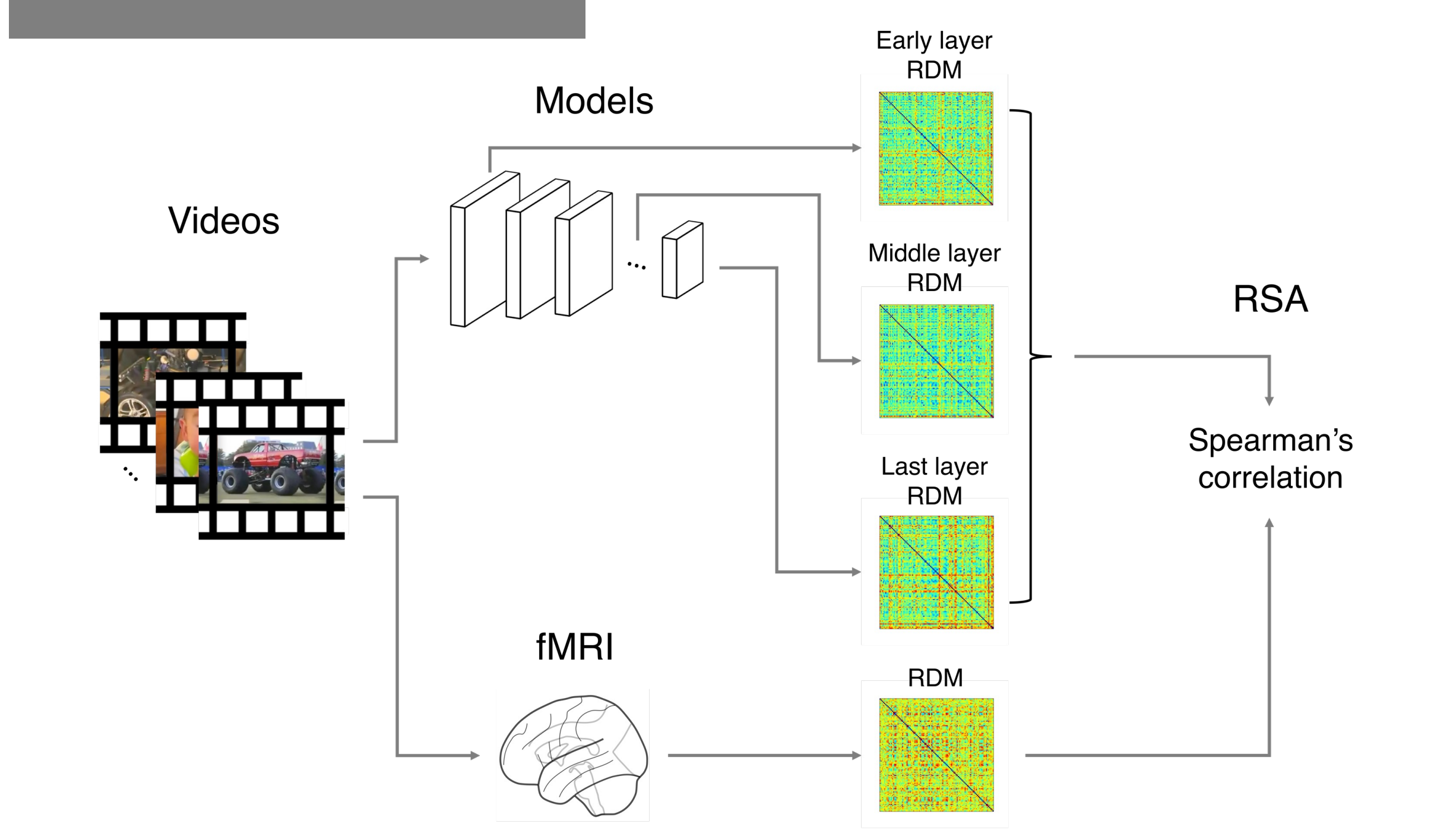
Hyewon Willow Han* [1,2,3], Ruchira Dhar* [1,4], Qingqing Yang* [1,5], Maryam Hoseini Behbahani[1],
María Alejandra Martínez Ortiz[1,6], Tolulope Samuel Oladele[1,7], Diana C Dima[2,3],
Hsin-Hung Li+ [5], Anders Søgaard+ [4], Yalda Mohsenzadeh+ [1,2,3]

Introduction

- The remarkable performance of transformer models in reasoning tasks and their widespread use have prompted much research on their alignment with brain activations.
- Key questions remain:
  - **Does alignment depend on input modality or training objective?**
  - **Is the alignment confined to modality-specific brain regions or extends to higher cognitive areas?**
- To explore this, we analyze representations of language, vision and vision-language transformer models and compare them with neural representations across multiple brain regions obtained during a visual task.

Methods

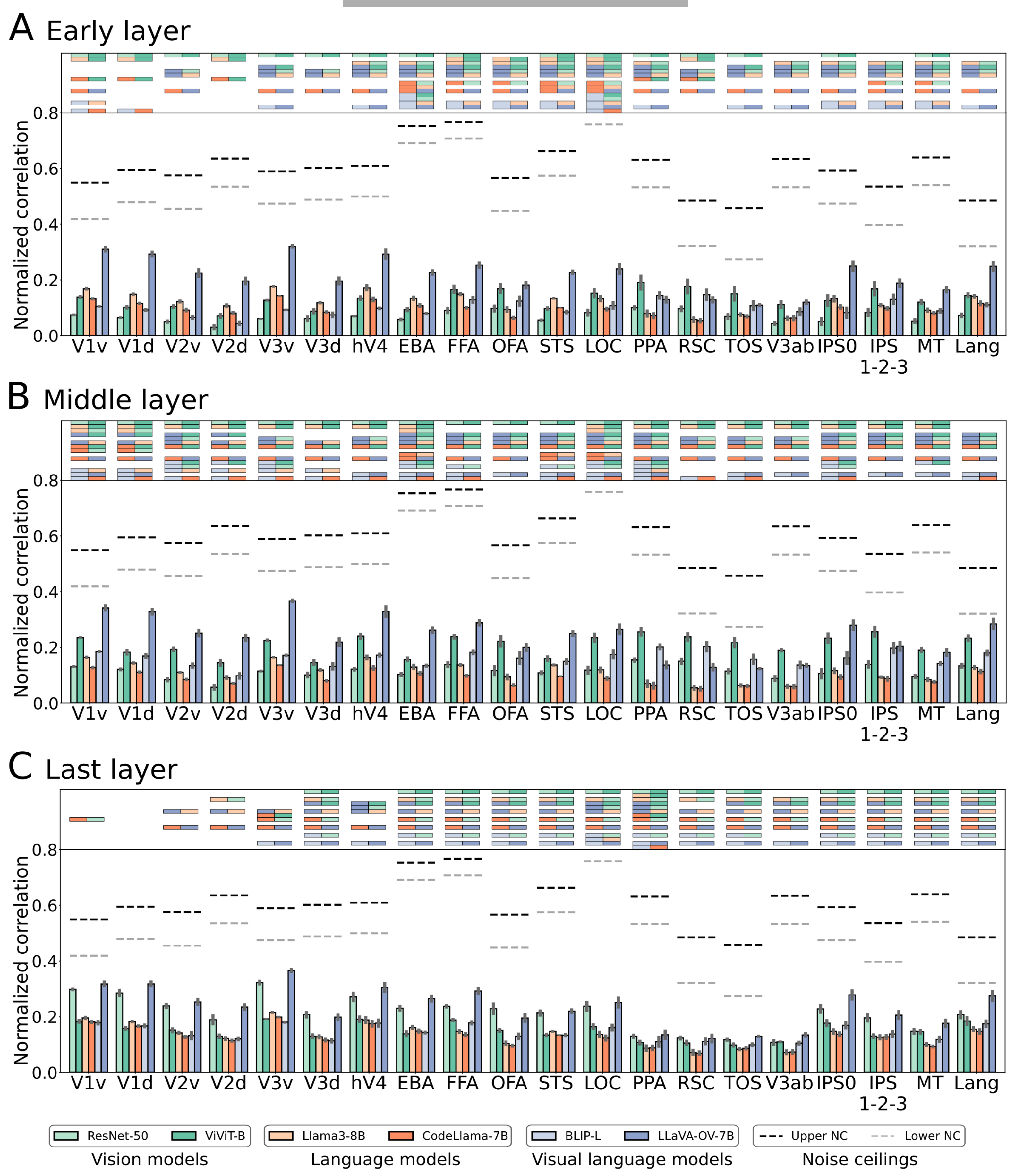




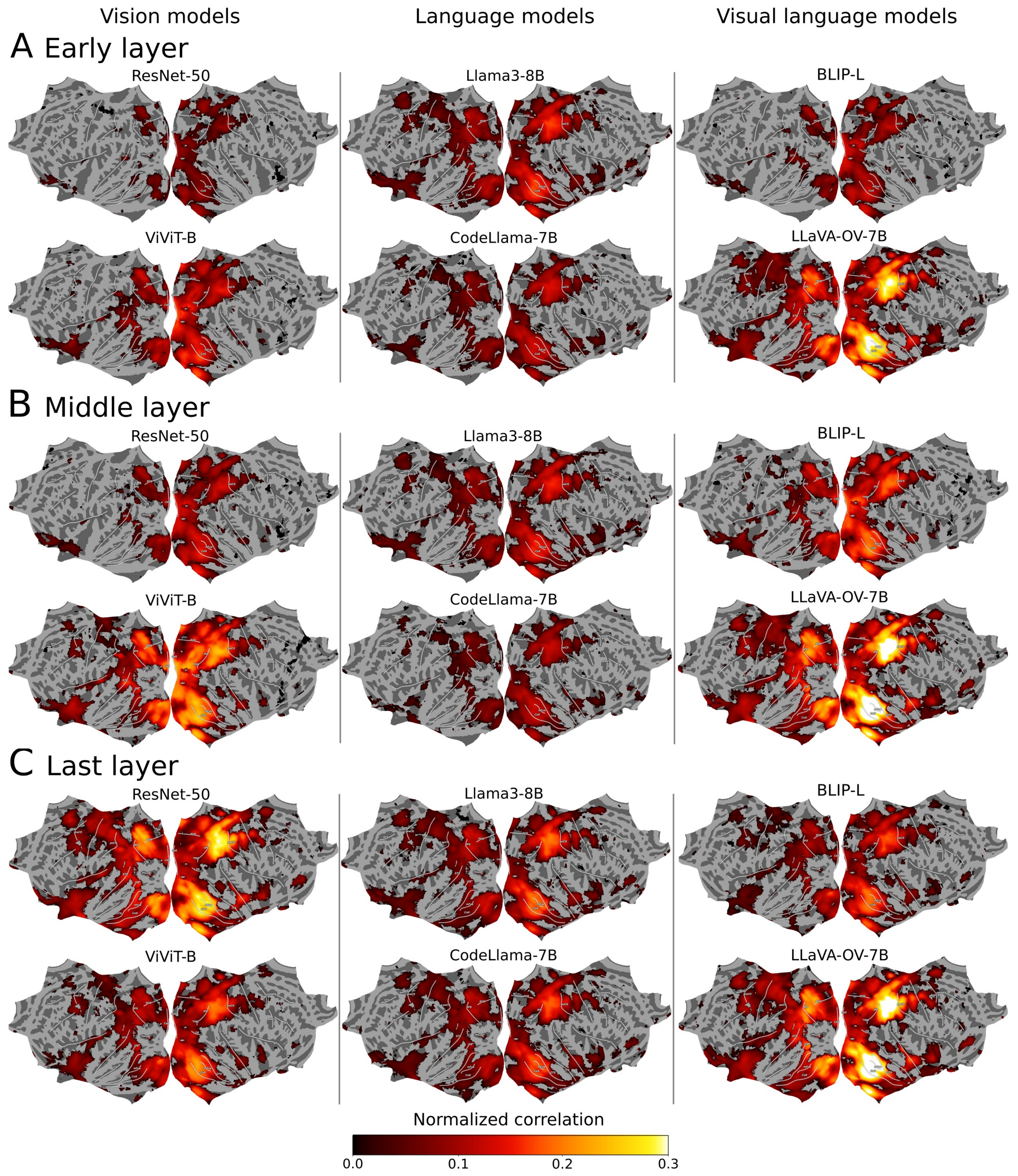
Model	Architecture	Modality	Training objective	Number of parameters	Used layers (early, middle, last)
ResNet-50	Convolutional Neural Network	Image	Classification	25M	1, 2, 4
ViViT-B	Transformer	Image	Classification	89M	3, 5, 12
Llama3-8B	Transformer	Natural Language	Predictive Processing	8B	8, 15, 32
CodeLlama-7B	Transformer	Natural Language, Code	Predictive Processing	7B	8, 15, 32
BLIP-L	Transformer	Image, Natural Language	Image Captioning	470M	6, 11, 24
LLaVA-OV-7B	Transformer	Video, Natural Language	Predictive Processing	7B	7, 13, 28

Results

ROI-based RSA



Searchlight RSA



Results Summary

- **Early layers**:
  - Transformer architectures generally outperform ResNet-50 in early visual regions.
  - BLIP-L aligns closely with LLaVA-OV-7B in early visual regions but diverges in other areas, emphasizing the role of training objectives in predictive processing.
  - LLaVA-OV-7B starts to align well with brain regions for mid- to high-level visual processing and cognitive control functions.
- **Mid layers**:
  - ResNet-50's mid-layer representations align better compared to its early layers, while transformer models maintain similar trends.
  - BLIP-L improves in early visual ROIs but remains outperformed by LLaVA-OV-7B, reinforcing the impact of predictive processing objectives.
  - ViViT-B and BLIP-L start to exhibit strong alignment with regions related to higher cognitive functions.
- **Last layers**:
  - ResNet-50 shows stronger performance in early visual areas compared to some transformers
  - ViViT-B and BLIP-L show reduced alignment in early regions, underlining the significance of predictive objectives for consistent representation.

Conclusion

- Our findings indicate that both the type of training data and the objectives used during training play a critical role in determining alignment, showing that models align with neural representations both within and beyond modality-specific regions.
- They also show that training modalities and objectives influence alignment quality across model layers, suggesting that using multimodal data combined with a predictive processing objective might provide more robust representational capabilities than alternative training objectives.
- These results emphasize the importance of multimodal training and predictive objectives in aligning models with human cognitive processes.

[1]Neuromatch Academy, [2]Western University, [3]Vector Institute,
[4]University of Copenhagen, [5]The Ohio State University,
[6]New York University, [7]University of Medical Sciences, Ondo

*These authors contributed equally to this work.
+Co-corresponding authors.

QR Code to Paper