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

Hyewon Willow Han[*1,2,3] hhan228@uwo.ca
Ruchira Dhar[*1,4] rudh@di.ku.dk
Qingqing Yang[*1,5] yang.6118@osu.edu
Maryam Hoseini Behbahani[1] maryam.hoseini2101@gmail.com
María Alejandra Martínez[1,6] mm13852@nyu.edu
Tolulope Oladele[1,7] toladele@unimed.edu.ng
Diana C. Dima[2,3] ddima@uwo.ca
Hsin-Hung Li[†5] li.14492@osu.edu
Anders Søgaard[†4] soegaard@di.ku.dk
Yalda Mohsenzadeh[†1,2,3] ymohsenz@uwo.ca

[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

Editors: Marco Fumero, Clementine Domine, Zorah Lähner, Donato Crisostomi, Luca Moschella, Kimberly Stachenfeld

Abstract

The remarkable performance of transformer models in both linguistic and real-world reasoning tasks coupled with their ubiquitous use has prompted much research on their alignment with brain activations. However, there remain some unanswered questions: what aspects of these models lead to representational alignment-the input modality or the training objective? Moreover, is the alignment limited to modality-specialized brain regions, or can representations align with brain regions involved in higher cognitive functions? To address this, we analyze the representations of different transformer architectures, including text-based and vision-based language models, and compare them with neural representations across multiple brain regions obtained during a visual processing task. Our findings reveal that both training data modality and training objective are important in determining alignment, and that models align with neural representations within and beyond the modality-specific regions. Additionally, the training modality and objectives seem to have an impact on alignment quality as we progress through the layers, suggesting that multimodal data along with a predictive processing objective may confer superior representational capabilities compared to other training objectives.

1 Introduction

The recent introduction of the transformer architecture [1], combined with a predictive processing training objective, has led to the rise of models that have achieved unprecedented performance in the

[*]These authors contributed equally to this work.
[†]Co-corresponding authors.

Proceedings of the II edition of the Workshop on Unifying Representations in Neural Models (UniReps 2024).

domain of natural language processing. Nowadays, larger variants of these language models, known as large language models (LLMs) or multimodal language models (MLMs), have been shown to have superior language understanding and generation capabilities [2, 3], structured understanding of language [4–7], and significant performance in broader cognitive domain tasks like general reasoning [8–12] and planning [13, 14]. This has also led to a proliferation of research with transformers within the neuroconnectionist research programme [15], focusing on alignment of their representations with brain activations [16–19], and there has been an increasing interest in several questions about their relation to human cognitive processes and how these models process and represent information compared to the human brain.

When it comes to humans, there has always been evidence of predictive processing playing an important role during language comprehension [20–22]. Predictive processing transformers today have shown superior capabilities in language processing - popular text-trained models like `Llama3-8B` [23] have shown robust language generation capabilities while those further trained on code like `CodeLlama-7B` [24] have also been shown to develop even advanced reasoning capabilities [25]. Moreover, the embeddings of transformer-based LMs have been shown to be robust at reflecting human judgements across language and vision inputs [26, 27]. This has led to the consideration of such transformer-based LMs as possible models of human language processing. However, while there is some research on determining pressures which impact such model alignment capabilities [28–30], there has been little work that specifically focuses on transformers and delineating design choices in them that improve alignment. Previous research has investigated the role of different architectures, task objectives and training diets on the alignment of biological and artificial systems [31, 32]. However, most of these studies have predominantly focused on visual processing and image-based tasks [33, 34], with relatively few exploring the multi-modal capabilities of artificial neural network models. On a related note, research has also shown that the human brain is highly modular: for example, linguistic and non-linguistic tasks are clearly separated from one another in the brain [35–38]. Recent work on the interface of LLMs and human language processing has also emphasized the need to separate language and general cognition [39, 40]. How valid is this domain specificity in the case of model representations?

These considerations lead us to our two central questions:

- Do input modality and training objective impact the representational alignment of transformers with the human brain?
- Can task or domain-specific representations from models align to brain regions with higher cognitive functions beyond modality-specialized regions?

To answer these, we consider the domain of visual processing tasks and compare representations from various deep generative models with brain activations across different regions. The stimuli and functional Magnetic Resonance Imaging (fMRI) data are taken from the BOLD Moments Dataset (BMD) [41], which contains fMRI data from 10 subjects viewing short natural videos. To study the impacts of training data modalities and objectives, we use six different types of models in our research: a convolutional neural network baseline model and five transformer architectures with varying input modalities and training objectives. Specifically, we use an image model (`ResNet-50`), a video model (`ViViTB`), a language model (`Llama3-8B`), a language model which was trained on programming languages (`CodeLlama-7B`), an image-language model (`BLIP-L`), and a video-language model (`LLaVA-OV-7B`). The code-trained language model was chosen because of the recent finding on its different behaviour on reasoning tasks compared to its natural text-based counterpart, `Llama-3-8B-Instruct` [11, 25]. We extract the models' hidden representations of the stimuli from their early, middle, and last layers, and apply representational similarity analysis (RSA) across 20 different human brain regions of interest (ROIs). Additionally, we apply searchlight RSA [42] to map where the model best reflects the local neural activation patterns.

Our results indicate that a combination of multimodal data and generalized predictive processing i.e. next-word prediction training objective is critical in improving the alignment with neural representations, the influence growing more pronounced as we hierarchically ascend model layers. This alignment also manifests for higher-level regions, highlighting the broad scope of representational convergence. This observation indicates that predictive processing, combined with multimodal data, may endow models with a more sophisticated and nuanced representational alignment capacity when compared to other training paradigms. Such results have considerable implications for research surrounding the cognitive capabilities of models and their ability to emulate human cognition.

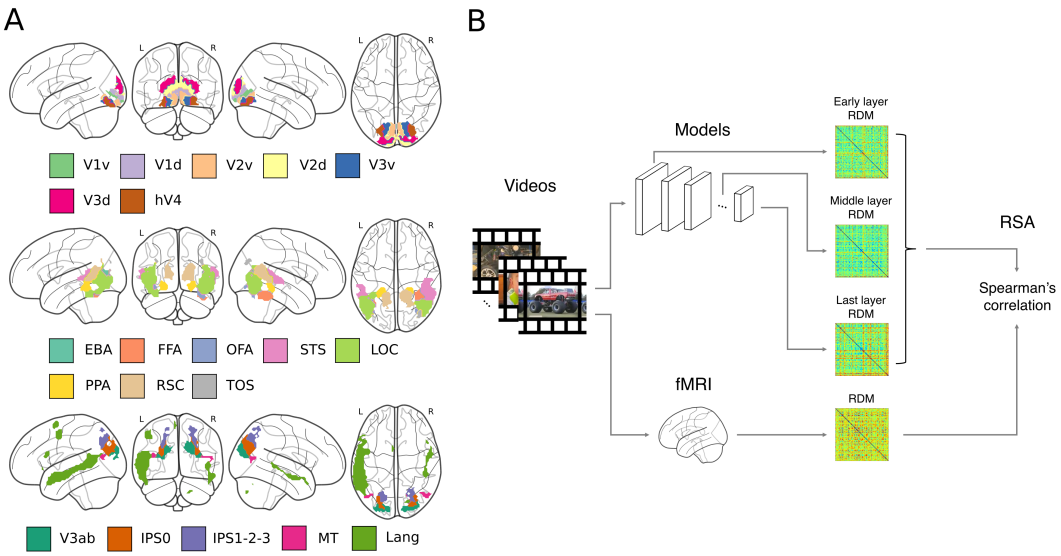

Figure 1: (**A**) Definitions of regions of interest (ROIs). Subject 1 was used for the visualization. (**B**) A schematic diagram of the comprehensive workflow for the analysis. For the visualization, `LLaVA-OV-7B` was used for model RDMs and V1v from Subject 1 was used for the fMRI RDM.

2 Methods

2.1 Stimuli

The stimuli used in our study consists of 102 short videos from the BOLD Moments Dataset (BMD) [41]. The videos presented to participants did not contain audio information or captions. In the scanner, subjects are instructed to fixate at the central fixation cross. Each video has a duration of 3 seconds and a frame rate between 15 and 30 frames per second, ensuring a diverse range of temporal resolutions. This test dataset is a carefully selected subset from the larger Memento10k dataset [43], which itself is derived from the extensive Moments in Time dataset [44] and its extended Multi-Moments in Time dataset [45]. The Memento10k dataset includes a broad spectrum of real-world activities and scenes, providing a rich context for evaluating vision-based models.

2.2 fMRI representations

BMD comprises data from 10 healthy subjects, with an average age of 27.01 years (SD = 3.96). Each subject participated in five fMRI sessions, including anatomical scans, functional localizer scans, and visual task fMRI sessions. During the main task, subjects viewed each video from the test set 10 times. They were instructed to maintain fixation on a central point throughout each video. Each video lasted 3 seconds and was followed by a 1-second interval.

We utilize 15 early and ventral visual ROIs defined from the localizer experiments conducted for each subject. These ROIs include early or mid-visual areas (V1v, V1d, V2v, V2d, V3v, V3d, hV4), which are critical for processing basic visual features. Additionally, body-selective regions (EBA), object-selective regions (LOC), face-selective regions (FFA, OFA, STS), and scene-selective regions (PPA, RSC, TOS) were included to capture specialized visual processing. The number of voxels for each ROI and each subject was capped at 1000, following the procedure described in the BMD paper, where the top 1000 ROIs were limited by masking each subject's FWE corrected t-contrast map with the corresponding binarized t-contrast probability map [41]. This approach allows for a detailed investigation into how different visual areas respond to the diverse stimuli presented in the videos.

Additionally, 3 dorsal visual stream areas defined from anatomical landmarks in BMD were adopted, including V3ab, IPS0, and IPS1-2-3 defined with a maximum probability map [46]. The middle temporal visual area (MT) was also extracted [47], as a part of the visual motion processing pathway [48, 49]. The language area (inferior frontal gyrus, inferior frontal gyrus orbital, middle frontal

Table 1: Detailed overview of model specifications

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

gyrus, posterior temporal region and anterior temporal region) was extracted for each individual with a probabilistic atlas, LanA [50]. In conclusion, 20 ROIs were included in the current study. The visualization of all 20 ROIs is shown in Figure 1A.

2.3 Model representations

To measure the impact of model design choices (input modality and training objective), we consider a suite of six models, including a baseline model: `ResNet-50` (convolutional neural net for image classification) [51], `ViViT-B` (video-vision transformer for classification) [52], `Llama3-8B` (a typical autoregressive language model) [23], `CodeLlama-7B` (LLAMA-variant further trained on code) [24], `BLIP-L` (a vision language model) [53], and `LLaVa-OneVison-7B` (a video language model) [54]. Refer to Table 1 for a more detailed overview of their architecture and training objectives. We consider 2 models with multimodal input i.e. image+language pretraining and 3 models with predictive processing i.e. next-word prediction objective where only `LLaVa-OneVison-7B` combines both i.e. it is a multimodal model with predictive processing objective.

For the text-based models (`Llama-3-8B-Instruct` and `CodeLlama-7B`), we use the caption data provided in the BMD. From a set of five given captions for each stimulus, we conducted a review of captions and chose the longest caption to ensure maximal information about the stimuli is preserved in the caption. For the image-based models (`ResNet-50`, `ViViT-B`, `BLIP-L`), we first extract 32 frames uniformly distributed across the duration of each video and then average the model representations across frames. For the video-based models (`ViViT-B` and `LLaVA-OV-7B`), we also input 32 frames uniformly distributed across the duration of each video.

For each model, we extracted three sets of representations while processing each stimulus: an early layer, an intermediate layer and the last layer. For the early layers, we extracted features from layers corresponding to a quarter of the total number of hidden state outputs for each model. The intermediate layers were extracted from the midpoint based on the number of hidden layers for each model, while the last layers were the last of the hidden layers in each model. The exact numbers of extracted layers are indicated in Table 1. For `ResNet-50`, the layers used in our analysis can be extracted by accessing each `Sequential` (PyTorch) layer, which is the block part of the ResNet model, and for all the other models, the layers can be extracted by accessing `hidden_states` (huggingface).

2.4 Representational dissimilarity matrices (RDM)

After extracting neural and model representations for the 102 video stimuli, we create RDMs to quantify how the stimuli were represented in each brain area and model. The RDMs enable the comparison of representation across systems. The dissimilarity between two representations X_i and X_j can be expressed as:

$$D(X_i, X_j) = 1 - \frac{\text{cov}(X_i, X_j)}{\sigma(X_i)\sigma(X_j)} \tag{1}$$

where $cov(X_i, X_j)$ denotes the covariance between the two representations, with $\sigma(X_i)$ being the standard deviation of X_i. This formula denotes a $1 - Pearson$ distance between the representations of video stimuli i and j.

The RDM is a symmetric nn matrix R where R_ij reflects the dissimilarity between the representations of stimulus i and stimulus j, resulting in a 102 by 102 matrix in the current study. Mathematically, the RDM is given by:

$$R_{ij} = D(X_i, X_j) \quad \forall i, j \in \{1, 2, \ldots, n\} \tag{2}$$

For the ROI-based analysis, we calculate RDMs for each subject and each ROI. For the whole brain searchlight analysis, RDMs are extracted within a sphere of radius 4 voxels centered at each voxel across the whole brain, measuring the local neural representation pattern.

2.5 Representational similarity analysis (RSA)

Representational Similarity Analysis (RSA) provides a common framework to quantitatively compare representational geometries across different modalities, such as computational models and neuroimaging data [42]. This approach has been particularly valuable in studying how both artificial neural network models and the human brain process complex, naturalistic stimuli like spoken or written language, images, and videos [55–57].

To assess the similarity of neural and model representations, we calculate the ($Spearman's\ \rho$) between RDMs:

$$\rho = \frac{\text{cov}(\text{rank}(\text{vec}(R_A)), \text{rank}(\text{vec}(R_B)))}{\sigma(\text{rank}(\text{vec}(R_A)))\sigma(\text{rank}(\text{vec}(R_B)))} \tag{3}$$

$$\rho = \text{Spearman}(\text{vec}(R_A), \text{vec}(R_B)) \tag{4}$$

We correlate neural RDMs with the RDMs of the early, middle and late layers of models to measure how well model representations capture brain responses to stimuli.

We compute the upper noise ceiling as subject-to-group RDM correlation and the lower noise ceiling as leave-one-out RDM correlation per ROI or searchlight[58]. For multivariate reliability, the RSA values are normalized by the upper noise ceiling for both analyses for each subject, as described by:

$$\rho_i^{\text{norm}} = \frac{\rho_i}{\rho_i^{\text{upper}}} \tag{5}$$

Afterwards, we average the normalized correlations across the 10 subjects at each ROI or searchlight. The final corrected RSA value $\rho_i^{corrected}$ is given by:

$$\rho_i^{\text{corrected}} = \frac{\rho_i^{\text{norm}}}{N} \tag{6}$$

To statistically assess the searchlight RSA results, we follow Lahner et al. [41] and compute a one-sample two-sided t-test at each searchlight testing whether correlations differed from 0. The resulting p-values are FDR-corrected across searchlights (assuming positive correlation, $q = 0.05$). For the ROI-based RSA, a one-way ANOVA was performed at each ROI level using noise-normalized correlation for all 6 models (Bonferroni corrected with $n = 20$ ROIs, $p < 0.05$) and a Tukey's Honestly Significant Difference (HSD) test was performed as a post-hoc test for significant ROIs (FEWR $= 0.05$). A schematic diagram of the RSA procedure is shown in Figure 1B.

3 Results

3.1 ROI-based RSA

We compare the similarity of model representations extracted from different layers to fMRI activations for corresponding stimuli across different brain regions, as shown in Figure 2.

For the representations in the early layers, we see that:

- Most transformer architectures perform better than the baseline `ResNet-50`, with `ViViT-B` showing highest alignment for the early visual regions (V1v, V1d, V2v, V2d) while `LLaVA-OV-7B` was more aligned across other ROIs. This could indicate that for higher-level brain regions, multiple input modalities lead to better alignment.
- The `BLIP-L` model shows near-comparable performance to `LLaVA-OV-7B` in the early visual ROIs, with the performance gap increasing in other brain regions. This suggests that over

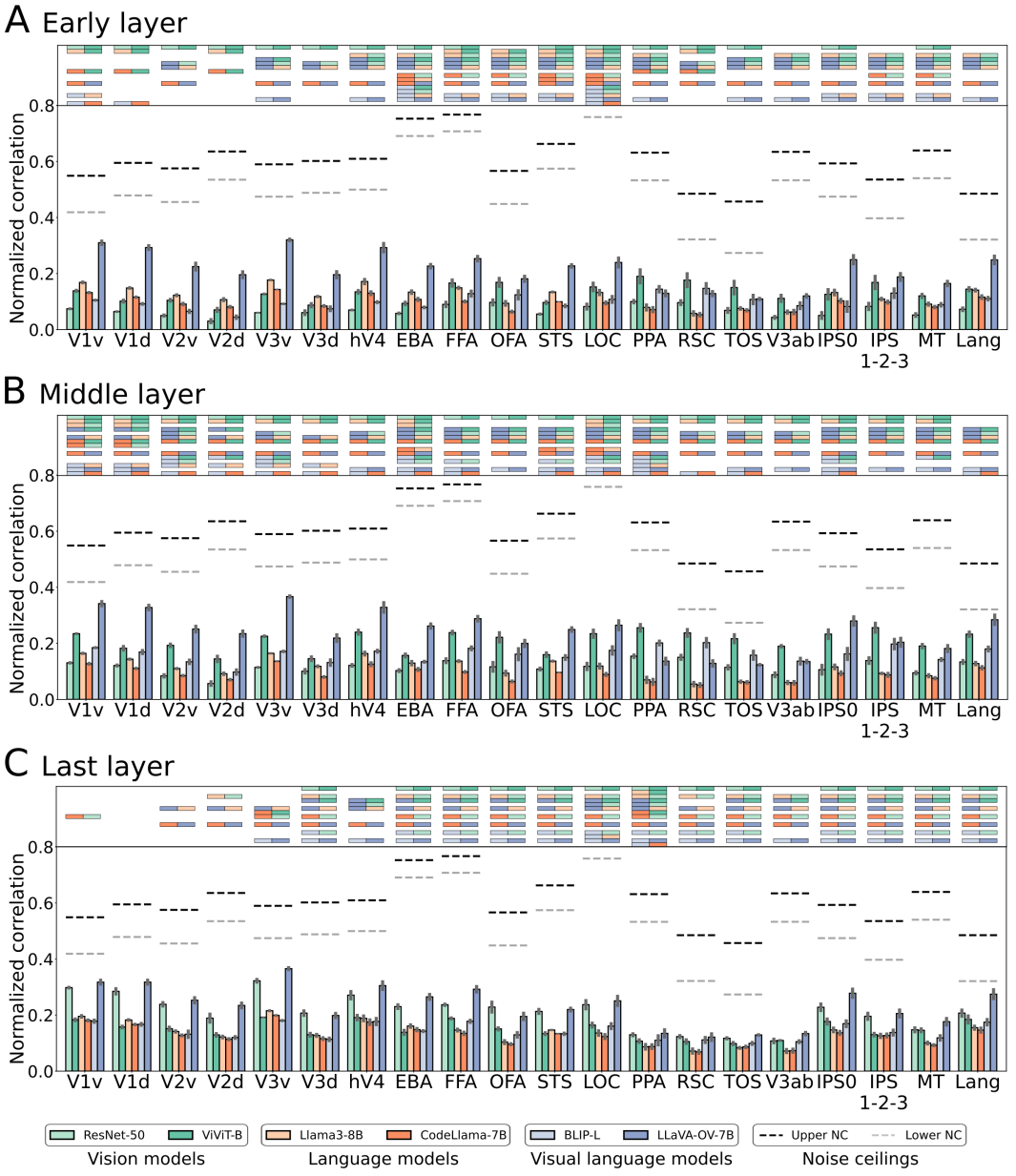

Figure 2: ROI-based RSA results. Noise-normalized Spearman's correlation coefficient values for (**A**) the early layer of each model, (**B**) the middle layer of each model, and (**C**) the last layer of each model. Error bars indicate the ± standard error. Noise normalization is done by using the upper bound of the noise ceiling for each ROI. Noise ceilings are shown for each ROI (lower bound: gray line; upper bound: black line). A one-way ANOVA test compared the noise-normalized correlation between all 6 models for each ROI (Bonferroni corrected, n = 20, $p < 0.05$). If significant, a Tukey's HSD test identified pairwise significance (FWER=0.05; significant pairs marked by dual paired color bars on top of each ROI plot).

and above input modality, the training objective also impacts performance, with improved alignment for predictive processing models.

- Text-based LMs (`Llama-3-8B-Instruct` and `CodeLlama-7B`) perform the worst in early and mid-level visual regions but show improvement in higher-level visual areas and dorsal, MT, Lang regions. This suggests that predictive processing leads to broader cognitive alignment but multiple input modalities tend to further improve performance (as seen by the better alignment of `LLaVA-OV-7B`).
- The `CodeLlama-7B` model mostly lags behind its natural language counterpart, `Llama-3-8B-Instruct`, suggesting that training with code, as opposed to only natural language, might lead to lower brain alignment even though their reasoning performance improves [24, 25]

For the mid-layer representations, we see that:

- The middle layer of the baseline `ResNet-50` achieves higher brain alignment compared to its early layer representations. The alignment of middle layer transformer representations is approximately similar to their early layers, with `ViViT-B` best modelling early visual regions (V1v, V1d, V2v, V2d, and V3d), and `LLaVA-OV-7B` showing high similarity to brain activations in other ROIs.
- For `BLIP-L`, the alignment trend is similar to that seen in its early layer, although the middle layer performs better in early visual ROIs. However, `LLaVA-OV-7B` still outperforms `BLIP-L`, supporting our conclusion that a predictive processing objective leads to better alignment.
- For the text-based LMs, the middle layer results are similar to the early layers, with `Llama-3-8B-Instruct` still outperforming `CodeLlama-7B`.

For the last layer representations of the model, we see some noteworthy patterns of alteration in alignment measures:

- The performance of `ResNet-50` is better compared to most other transformers in the early visual areas (which replicates some earlier findings [32]) but `LLaVA-OV-7B` still remains the best-performing model across all ROIs. Surprisingly, however, the alignment for `ViViT-B` and `BLIP-L` drops for early visual ROIs. This could indicate the importance of a predictive processing objective for retaining representational nuances across model layers.
- The performance of text-based LMs also remains similar or shows slight improvements across all ROIs, yet`LLaVA-OV-7B` still outperforms the text-based models. This highlights the superiority of multimodal models even within the predictive processing model class.

Moreover, we find that the representations from all our 6 models are similarly well-aligned not only to modality-specific regions (early visual areas or the language network), but also to regions involved in higher cognitive functions like IPSO or IPS1-2-3. This further strengthens the view that model representations also capture information in regions beyond low-level sensory processing areas.

3.2 Searchlight RSA

The whole-brain searchlight analysis is designed to provide a more detailed understanding of "where" in the human brain specific local response patterns exhibit similarity to the way the model encodes and represents the videos. By systematically evaluating neural activity across the entire brain, this method enables the identification of precise regions that align with the model's representational structure. The findings from the ROI-based (Region of Interest) analysis, which focused on predefined brain areas, were further validated and strengthened by the searchlight approach, offering a more comprehensive, data-driven confirmation of the model-brain alignment across broader cortical regions, as shown in Figure 3.

For the representations in the early layers, we see that:

- Image-based models `ViViT-B` and `RestNet-50` exhibit diverse alignment across different brain regions while `ViViT-B` shows better alignment for early to middle visual areas.

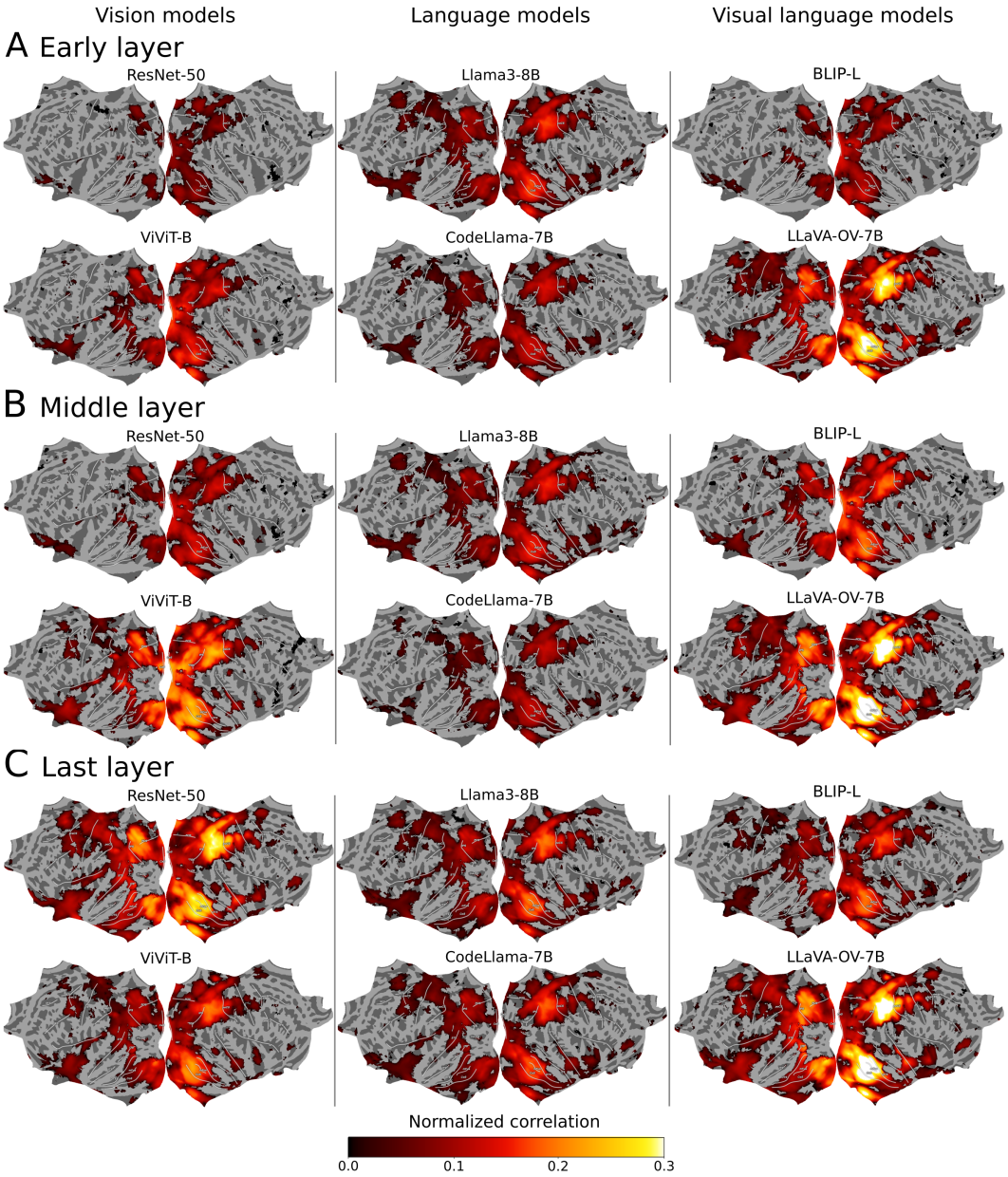

Figure 3: Searchlight RSA results. Noise-normalized Spearman's correlation coefficient values for (**A**) the early layer of each model, (**B**) the middle layer of each model, and (**C**) the last layer of each model. From left to right, the columns show non-language, language-only, and multimodal models. Noise normalization was done by using the upper bound of the noise ceiling at each searchlight. FDR correction ($q = 0.05$) was performed across searchlights.

- For video+text based model `LLaVA-OV-7B`, the alignment performance is the best among all six models. Though this model does not reflect the activation pattern of early visual areas, it aligns well with brain regions for middle to high-level visual processing and cognitive control functions.

This shows that while transformer architectures with predictive processing are more brain-aligned, enhancement with visual input modalities can further improve the alignment of representations.

In case of the middle layer representations, the trends seems to indicate that training with image data, image-based model ViViT-B and image+text-based model BLIP-L show stronger alignment compared to the early layer. Meanwhile, they exhibit strong alignment with superior parietal areas, which are related to higher cognitive functions such as attention modulation and multimodal sensory integration.

For the last layer representations of models, the trends are as follows:

- Similar to ROI-based analysis, the alignment for image-based model ViViT-B and image+text-based model BLIP-L drops for all the voxels that their the middle layer reflect, especially for early visual ROIs.
- Baseline image-based model RestNet-50 gained stronger alignment overall compared to the middle layer, especially the early visual ROIs and abstract information selective ROIs.
- The video-text based model LLaVA-OV-7B still performed the best in brain alignment, remaining the strong mapping to the brain areas that reflects by middle layer.

Similar to the ROI-based analysis, the searchlight analysis reveals that the transformer architecture yields the highest alignments when combined with a next-word prediction objective and multimodal training data.

4 Discussion

The findings of this study offer valuable insights into the factors that shape the alignment of transformer-based models with neural activity in the human brain. Our results emphasize the significance of architectural design, training strategies, and task modalities in influencing the representational overlap between artificial models and biological systems.

Impact of Training Data and Multimodal Learning While transformers with predictive processing objectives (Llama-3-8B-Instruct, CodeLlama-7B and LLaVA-OV-7B) demonstrate strong representational alignment with human brain regions, our findings suggest that the training data also plays a critical role in improving this alignment. The CodeLlama-7B model mostly lags behind its natural language counterpart, Llama-3-8B-Instruct, suggesting that training with code, as opposed to only natural language, might narrow model's ability to capture natural language patterns and features as effectively. Additionally, models trained on multiple input modalities, such as both vision and language, tend to exhibit stronger and more generalized alignment with brain activity (LLaVA-OV-7B). This multimodal training likely enables models to capture richer, more integrated representations that better reflect the multifaceted nature of human cognition, where various sensory and cognitive inputs are processed simultaneously.

The human brain continuously integrates information from different sources—visual, auditory, linguistic, and more—during perception and decision-making. By training transformers on both vision and language, for example, the models are able to capture complex interactions between these modalities, resulting in a more holistic and cognitively aligned representation that resonate more with how the brain processes diverse inputs [59]. This is particularly evident in non-linguistic brain regions where models trained solely on language show weaker alignment, whereas multimodal-trained models achieve a more robust alignment across diverse cognitive networks. This suggests that incorporating multimodal data into training protocols enhances representational alignment capabilities.

Impact of Training Objective The training objective seems to play a crucial role in determining brain alignment. Transformer-based models like ViViT-B and BLIP-L show worsening alignment with early visual areas as we pass through model layers, which can be attributed to their alternate training objectives such as classification or image captioning. On the other hand, the models with next-word prediction objectives such as Llama-3-8B-Instruct, CodeLlama-7B and LLaVA-OV-7B all improve and retain alignment even at the last layer representations. This suggests that predictive processing objectives might better reflect cognitive mechanisms in the brain [60, 61], also evidenced by studies of the human visual system's higher-level areas providing predictive signals into lower regions to "explain away" things [62–65]. These results could also count as supporting evidence of the human visual system as a predictive machine. Moreover, the increasing alignment observed in higher layers of the models may indicate that more abstract, high-level representations in the brain and models converge when predictive objectives are used.

These findings highlight the potential of predictive processing frameworks not only to improve the performance of artificial models but also to enhance their ability to simulate human cognitive processes. The success of these models in aligning with human brain activity may encourage further research into training paradigms that prioritize prediction as a core objective.

Broad-scope Cognitive Alignment Our findings demonstrate that the representations from all six models not only maintain alignment trends with modality-specific regions (such as early visual areas and the language network) but also extend to regions associated with higher-order cognitive functions, including IPSO, IPS1-2-3, and superior parietal areas, which aligns with previous research [66]. This supports the notion that model representations are capable of aligning with brain regions beyond low-level sensory processing, offering insights into more complex neural dynamics.

Moreover, this broad-scope alignment suggests that the models are not confined to superficial sensory features but are capable of capturing abstract cognitive processes, suggesting that they might share an abstract representation space beyond respective modalities.

Conclusion Our study raises critical questions about the nature of representational capabilities in transformer LMs- particularly as a function of their training data and objectives. The results of this study have broader implications for both neuroscience and artificial intelligence research. In the field of AI, the findings provide valuable insights into how architectural and training choices impact a model's ability to mimic human cognitive processes. Our results suggest that the combination of multimodal training and predictive processing objectives may be particularly effective in developing models that align with human neural patterns, opening new possibilities for creating more cognitively aligned artificial systems. From a neuroscientific perspective, transformer architectures may offer new opportunities to explore how the brain processes information across various domains. Their ability to align with low to high-level cognitive brain regions suggests that these models can be useful tools for studying the neural basis of cognition more broadly.

5 Limitations

Despite the promising findings, our study has a few limitations that must be considered when interpreting the results. First, the analysis primarily focuses on visual processing and language comprehension tasks, limiting the generalisability of our results to other cognitive domains. Cognitive processes such as memory, attention, and abstract reasoning are not directly explored, leaving open the question of how well transformer-based models align with brain regions involved in these tasks. Future studies should incorporate a broader range of cognitive tasks to determine whether the representational alignment observed here extends to other domains of cognition.

Secondly, our study only tests a selection of transformer models. To validate our findings and ensure the robustness of representational alignment, future research should include a wider range of models to assess how generalizable our results are and whether certain model architectures or training paradigms consistently produce stronger alignment with neural activity across multiple cognitive domains. This approach would provide a deeper understanding of which specific model features contribute most to effective alignment with the human brain.

Finally, while we demonstrate an alignment of models with brain regions, it is important to recognize that alignment does not necessarily imply equivalence in cognitive mechanisms. The models we analyzed are optimized for prediction and generation in artificial tasks, and their internal representations may not map directly onto biological processes in a straightforward way. Thus, further work is needed to establish a deeper understanding of how these models function as proxies for brain activity.

Acknowledgments and Disclosure of Funding

The authors would like to express their sincere gratitude to Neuromatch Academy for providing an outstanding platform for learning, collaboration, and research. The online course and project facilitated by Neuromatch Academy not only brought the authors together but also provided valuable opportunities for presenting earlier versions of this work and receiving critical feedback from the broader community. We would also like to thank our mentors, colleagues, and peers for their insightful comments and support throughout the research process.

The work was supported by a Vector Institute Research Grant and an Natural Sciences and Engineering Research Council of Canada Discovery Grant to Y.M.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need.(nips), 2017. *arXiv preprint arXiv:1706.03762*, 10: S0140525X16001837, 2017.

[2] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[4] Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew E Peters, and Noah A Smith. Linguistic knowledge and transferability of contextual representations. *arXiv preprint arXiv:1903.08855*, 2019.

[5] Tal Linzen and Marco Baroni. Syntactic structure from deep learning. *Annual Review of Linguistics*, 7(1): 195–212, 2021.

[6] Vipula Rawte, Kaushik Roy, Megha Chakraborty, Manas Gaur, Keyur Faldu, Prashant Kikani, Hemang Akbari, Amit Sheth, et al. Tdlr: Top (semantic)-down (syntactic) language representation. In *Attention Workshop, 36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, 2022.

[7] Ben Ambridge and Liam Blything. Large language models are better than theoretical linguists at theoretical linguistics. *Theoretical Linguistics*, 50(1-2):33–48, 2024.

[8] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[9] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[10] Taylor Webb, Keith J Holyoak, and Hongjing Lu. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541, 2023.

[11] Ruchira Dhar and Anders Søgaard. From words to worlds: Compositionality for cognitive architectures. *arXiv preprint arXiv:2407.13419*, 2024.

[12] Abulhair Saparov, Richard Yuanzhe Pang, Vishakh Padmakumar, Nitish Joshi, Mehran Kazemi, Najoung Kim, and He He. Testing the general deductive reasoning capacity of large language models using ood examples. *Advances in Neural Information Processing Systems*, 36, 2024.

[13] Tom Silver, Varun Hariprasad, Reece S Shuttleworth, Nishanth Kumar, Tomás Lozano-Pérez, and Leslie Pack Kaelbling. Pddl planning with pretrained large language models. In *NeurIPS 2022 foundation models for decision making workshop*, 2022.

[14] Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. Layoutgpt: Compositional visual planning and generation with large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

[15] Adrien Doerig, Rowan P Sommers, Katja Seeliger, Blake Richards, Jenann Ismael, Grace W Lindsay, Konrad P Kording, Talia Konkle, Marcel AJ Van Gerven, Nikolaus Kriegeskorte, et al. The neuroconnectionist research programme. *Nature Reviews Neuroscience*, 24(7):431–450, 2023.

[16] Mycal Tucker and Greta Tuckute. Increasing brain-llm alignment via information-theoretic compression. In *UniReps: the First Workshop on Unifying Representations in Neural Models*, 2023.

[17] Adrien Doerig, Tim C Kietzmann, Emily Allen, Yihan Wu, Thomas Naselaris, Kendrick Kay, and Ian Charest. Visual representations in the human brain are aligned with large language models, 2022.

[18] Jiaang Li, Antonia Karamolegkou, Yova Kementchedjhieva, Mostafa Abdou, Sune Lehmann, and Anders Søgaard. Structural similarities between language models and neural response measurements. *arXiv preprint arXiv:2306.01930*, 2023.

[19] Khai Loong Aw, Syrielle Montariol, Badr AlKhamissi, Martin Schrimpf, and Antoine Bosselut. Instruction-tuning aligns llms to the human brain. In *First Conference on Language Modeling*, 2023.

[20] Nathaniel J Smith and Roger Levy. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319, 2013.

[21] Roel M Willems, Stefan L Frank, Annabel D Nijhof, Peter Hagoort, and Antal Van den Bosch. Prediction during natural language comprehension. *Cerebral cortex*, 26(6):2506–2516, 2016.

[22] Micha Heilbron, Kristijan Armeni, Jan-Mathijs Schoffelen, Peter Hagoort, and Floris P De Lange. A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, 119(32):e2201968119, 2022.

[23] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models, 2024. URL `https://arxiv.org/abs/2407.21783`.

[24] Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, et al. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*, 2023. doi: 10.48550/arXiv.2308.12950.

[25] YINGWEI MA, Yue Liu, Yue Yu, Yuanliang Zhang, Yu Jiang, Changjian Wang, and Shanshan Li. At which training stage does code data help LLMs reasoning? In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=KIPJKST4gw`.

[26] Diana C Dima, Sugitha Janarthanan, Jody C Culham, and Yalda Mohsenzadeh. Shared representations of human actions across vision and language. *Neuropsychologia*, 202:108962, 2024.

[27] Changde Du, Kaicheng Fu, Bincheng Wen, Yi Sun, Jie Peng, Wei Wei, Ying Gao, Shengpei Wang, Chuncheng Zhang, Jinpeng Li, et al. Human-like object concept representations emerge naturally in multimodal large language models. *arXiv preprint arXiv:2407.01067*, 2024.

[28] Florentin Guth and Brice Ménard. On the universality of neural encodings in cnns. In *ICLR 2024 Workshop on Representational Alignment*, 2024.

[29] Andrew Ligeralde, Yilun Kuang, Thomas Edward Yerxa, Miah N Pitcher, Marla Feller, and SueYeon Chung. Unsupervised learning on spontaneous retinal activity leads to efficient neural representation geometry. In *Proceedings of UniReps: the First Workshop on Unifying Representations in Neural Models*, pages 194–208. PMLR, 2024.

[30] Zorah Lähner and Michael Moeller. On the direct alignment of latent spaces. In *Proceedings of UniReps: the First Workshop on Unifying Representations in Neural Models*, pages 158–169. PMLR, 2024.

[31] Colin Conwell, Jacob S Prince, George A Alvarez, and Talia Konkle. What can 5.17 billion regression fits tell us about artificial models of the human visual system? In *SVRHM 2021 Workshop@ NeurIPS*, 2021.

[32] Colin Conwell, Jacob S Prince, Kendrick N Kay, George A Alvarez, and Talia Konkle. What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *BioRxiv*, pages 2022–03, 2022.

[33] Radoslaw Martin Cichy, Kshitij Dwivedi, Benjamin Lahner, Alex Lascelles, Polina Iamshchinina, Monika Graumann, Alex Andonian, NAR Murty, K Kay, Gemma Roig, et al. The algonauts project 2021 challenge: How the human brain makes sense of a world in motion. *arXiv preprint arXiv:2104.13714*, 2021.

[34] Alessandro T Gifford, Benjamin Lahner, Sari Saba-Sadiya, Martina G Vilas, Alex Lascelles, Aude Oliva, Kendrick Kay, Gemma Roig, and Radoslaw M Cichy. The algonauts project 2023 challenge: How the human brain makes sense of natural scenes. *arXiv preprint arXiv:2301.03198*, 2023.

[35] Evelina Fedorenko, Michael K Behr, and Nancy Kanwisher. Functional specificity for high-level linguistic processing in the human brain. *Proceedings of the National Academy of Sciences*, 108(39):16428–16433, 2011.

[36] Martin M Monti, Lawrence M Parsons, and Daniel N Osherson. Thought beyond language: neural dissociation of algebra and natural language. *Psychological science*, 23(8):914–922, 2012.

[37] John P Coetzee, Micah A Johnson, Youngzie Lee, Allan D Wu, Marco Iacoboni, and Martin M Monti. Dissociating language and thought in human reasoning. *Brain Sciences*, 13(1):67, 2022.

[38] Cory Shain, Alexander Paunov, Xuanyi Chen, Benjamin Lipkin, and Evelina Fedorenko. No evidence of theory of mind reasoning in the human language network. *Cerebral Cortex*, 33(10):6299–6319, 2023.

[39] Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 2024.

[40] Greta Tuckute, Nancy Kanwisher, and Evelina Fedorenko. Language in brains, minds, and machines. *Annual Review of Neuroscience*, 47, 2024.

[41] Benjamin Lahner, Kshitij Dwivedi, Polina Iamshchinina, Monika Graumann, Alex Lascelles, Gemma Roig, Alessandro Thomas Gifford, Bowen Pan, SouYoung Jin, N Apurva Ratan Murty, et al. Modeling short visual events through the bold moments video fmri dataset and metadata. *Nature Communications*, 15(1):6241, 2024.

[42] Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249, 2008. doi: 10.3389/neuro.06.004.2008.

[43] Anelise Newman, Camilo Fosco, Vincent Casser, Allen Lee, Barry McNamara, and Aude Oliva. Multimodal memorability: Modeling effects of semantics and decay on video memorability. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 223–240. Springer, 2020. doi: https://doi.org/10.1007/978-3-030-58517-4_14.

[44] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):502–508, 2019. doi: https://doi.org/10.1109/TPAMI.2019.2901464.

[45] Mathew Monfort, Bowen Pan, Kandan Ramakrishnan, Alex Andonian, Barry A McNamara, Alex Lascelles, Quanfu Fan, Dan Gutfreund, Rogério Schmidt Feris, and Aude Oliva. Multi-moments in time: Learning and interpreting models for multi-action video understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9434–9445, 2021. doi: https://doi.org/10.1109/TPAMI.2021.3126682.

[46] Liang Wang, Ryan EB Mruczek, Michael J Arcaro, and Sabine Kastner. Probabilistic maps of visual topography in human cortex. *Cerebral cortex*, 25(10):3911–3931, 2015. doi: 10.1093/cercor/bhu277.

[47] Matthew F Glasser, Timothy S Coalson, Emma C Robinson, Carl D Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F Beckmann, Mark Jenkinson, et al. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016. doi: 10.1038/nature18933.

[48] Richard T Born and David C Bradley. Structure and function of visual area mt. *Annu. Rev. Neurosci.*, 28 (1):157–189, 2005.

[49] Shinji Nishimoto and Jack L. Gallant. A three-dimensional spatiotemporal receptive field model explains responses of area mt neurons to naturalistic movies. *Journal of Neuroscience*, 31(41):14551–14564, 2011. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.6801-10.2011. URL `https://www.jneurosci.org/content/31/41/14551`.

[50] Benjamin Lipkin, Greta Tuckute, Josef Affourtit, Hannah Small, Zachary Mineroff, Hope Kean, Olessia Jouravlev, Lara Rakocevic, Brianna Pritchett, Matthew Siegelman, et al. LanA (language atlas): A probabilistic atlas for the language network based on fmri data from >800 individuals. *bioRxiv*, 2022. doi: 10.1101/2022.03.06.483177. URL `https://www.biorxiv.org/content/early/2022/03/07/2022.03.06.483177`.

[51] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[52] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021.

[53] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. URL `https://arxiv.org/abs/2201.12086`.

[54] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[55] Aria Wang, Michael Tarr, and Leila Wehbe. Neural taskonomy: Inferring the similarity of task-derived representations from brain activity. *Advances in neural information processing systems*, 32, 2019.

[56] Alex H Williams, Erin Kunz, Simon Kornblith, and Scott Linderman. Generalized shape metrics on neural representations. *Advances in Neural Information Processing Systems*, 34:4738–4750, 2021.

[57] Lukas Muttenthaler, Lorenz Linhardt, Jonas Dippel, Robert A Vandermeulen, Katherine Hermann, Andrew Lampinen, and Simon Kornblith. Improving neural network representations using human similarity judgments. *Advances in Neural Information Processing Systems*, 36, 2024.

[58] Hamed Nili, Cai Wingfield, Alexander Walther, Li Su, William Marslen-Wilson, and Nikolaus Kriegeskorte. A toolbox for representational similarity analysis. *PLoS computational biology*, 10(4):e1003553, 2014.

[59] Jerry Tang, Meng Du, Vy Vo, Vasudev Lal, and Alexander Huth. Brain encoding models based on multimodal transformers can transfer across language and vision. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 29654–29666. Curran Associates, Inc., 2023. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/5ebbbac62b968254093023f1c95015d3-Paper-Conference.pdf`.

[60] Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. Shared computational principles for language processing in humans and deep language models. *Nature neuroscience*, 25(3):369–380, 2022. doi: https://doi.org/10.1038/s41593-022-01026-4.

[61] Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature human behaviour*, 7(3):430–441, 2023. doi: https://doi.org/10.1038/s41562-022-01516-2.

[62] Andy Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3):181–204, 2013. doi: https://doi.org/10.1017/S0140525X12000477.

[63] Rajesh PN Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79–87, 1999. doi: https://doi.org/10.1038/4580.

[64] Karl Friston. A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences*, 360(1456):815–836, 2005. doi: https://doi.org/10.1098/rstb.2005.1622.

[65] Peter Kok and Floris P de Lange. Predictive coding in sensory cortex. *An introduction to model-based cognitive neuroscience*, pages 221–244, 2015. doi: https://doi.org/10.1007/978-1-4939-2236-9_11.

[66] Bhavin Choksi, Milad Mozafari, Rufin Vanrullen, and Leila Reddy. Multimodal neural networks better explain multivoxel patterns in the hippocampus. *Neural Networks*, 154:538–542, 2022.