Dissecting the Neuro-Cognitive Dynamics of Vision-Language Models

Jingyang Xue[1], **Runze Xia**[2] **& Peijian Zhang**[*1]
[1]College of Computer Science and Technology, Qingdao University, Qingdao, China
[2]College of Artificial Intelligence, Nanjing University of Aeronautics and Astronautics, Nanjing, China

Abstract

Vision-Language Models (VLMs) have achieved remarkable success in multimodal reasoning, yet the extent to which their internal representations mirror human neural dynamics remains an open scientific question. In this study, we propose the Component-Specific Residual Disentanglement (CSRD) framework to systematically evaluate the alignment between 17 state-of-the-art VLMs and human fMRI activity. Unlike traditional layer-wise approaches, CSRD disentangles the contributions of Attention and Feed-Forward Networks (FFNs) by selectively retaining or stripping residual connections, thereby isolating information integration from functional transformation. Leveraging the Natural Scenes Dataset (NSD), our analysis reveals a hierarchical "ascent" in brain-model alignment, but with a critical functional divergence in deep layers: Attention modules act as "cognitive hubs" maintaining high alignment with the human visual cortex, whereas FFNs progressively decouple from biological signals to prioritize task-specific predictions. Furthermore, we find that instruction-tuned generative architectures significantly outperform contrastive baselines (e.g., CLIP) and scaling-law-based expectations. These findings suggest that the generative objective fosters more biologically plausible representational structures, providing new theoretical guidance for designing brain-inspired multimodal systems.

Keywords: Vision-Language Models; Brain Alignment; fMRI; Representational Similarity Analysis; Transformer Components

Introduction

Recent advances in artificial intelligence have been significantly driven by Vision-Language Models (VLMs), which integrate visual and textual understanding through large-scale pre-training on image-text pairs followed by instruction tuning (Kaplan et al., 2020). Representative architectures such as CLIP (Radford et al., 2021) and LLaVA (Liu et al., 2023) not only achieve excellent performance on classic tasks like image captioning and retrieval, but also demonstrate emergent cross-modal reasoning capabilities that begin to mirror aspects of human perception and cognition (J. Bai et al., 2023; Bubeck et al., 2023; Chen et al., 2024; Wu et al., 2024)

Despite their empirical success, the internal mechanisms by which VLMs align visual and linguistic representations remain poorly understood. While some interpretability studies probe intermediate hidden states (Alain & Bengio, 2016; Belinkov, 2021), these models largely function as opaque systems. In contrast, cognitive neuroscience has established a detailed account of how the human brain integrates vision and language—particularly through the interaction between the ventral visual stream and high-level language areas during multimodal processing. This raises a compelling scientific question: *To what extent do modern VLMs recapitulate the neural dynamics observed in the human brain when representing real-world scenes?* Addressing this question offers not only a biologically grounded lens for interpreting model behavior, but also principled guidance for designing more cognitively plausible artificial systems.

However, current multimodal systems, while powerful, often operate as 'black boxes' where the specific contribution of architectural sub-components to biological alignment remains obscured. Dissecting these internal dynamics is crucial for distinguishing between superficial correlation and genuine mechanistic convergence, ensuring that our models are not just mimicking human behavior but are underpinned by similar computational principles.

Prior work has explored brain-model alignment primarily in unimodal settings. For instance, activations in convolutional networks correlate with fMRI responses along the ventral visual pathway (Guclu & van Gerven, 2015; Yamins et al., 2014), while language model representations align with activity in language-selective cortical regions (Schrimpf et al., 2021). However, extending such analyses to multimodal models faces two key challenges. First, most studies treat Transformer layers as homogeneous computational blocks (Caucheteux et al., 2022), overlooking the distinct functional roles of their internal components—Self-Attention (which routes contextual information) and Feed-Forward Networks (FFNs, which store and transform feature knowledge) (Clark et al., 2019; Geva et al., 2021). Moreover, it remains unclear how these sub-modules differentially engage with neural signals across modalities during cross-modal integration, as existing alignment frameworks lack the granularity to disentangle their contributions.

To address these challenges, we propose the *Component-Specific Residual Disentanglement* (CSRD) protocol—a fine-grained analysis framework that separately evaluates the brain alignment of Attention and FFN pathways within VLMs. Specifically, CSRD preserves residual connections for Attention to capture global integration effects, while isolating raw FFN outputs to assess their independent representational transformations. Using the Natural Scenes Dataset (NSD) (Allen et al., 2022)—a large-scale fMRI benchmark of human visual and linguistic responses to natural images—we apply

* Corresponding author:peijianzh@126.com

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

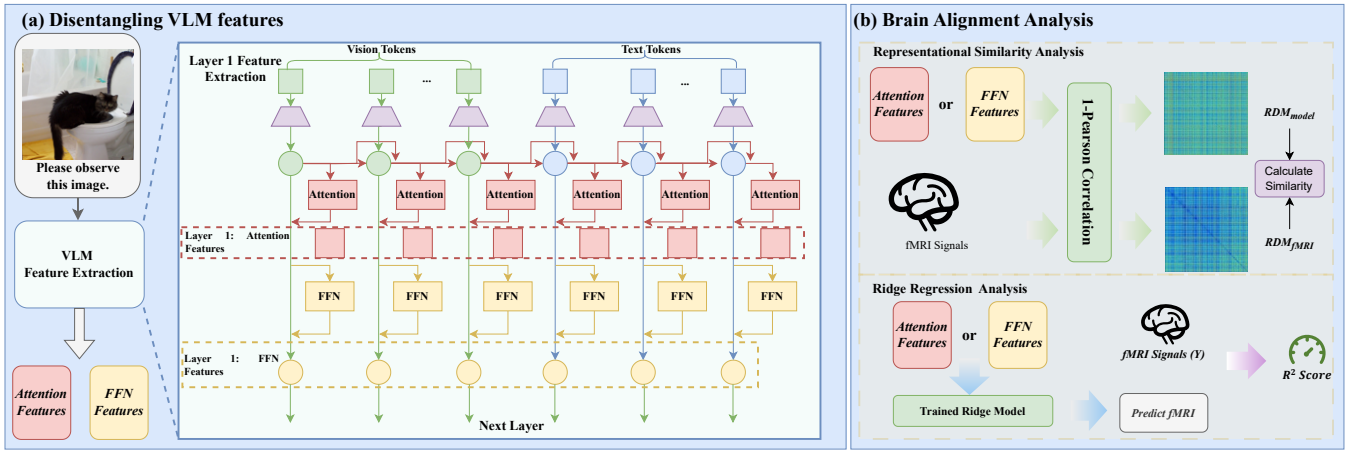

Figure 1: Overview of the CSRD framework. (a) Feature Disentanglement: Illustration of extracting representations from Attention (with residual retention) and FFN (with residual stripping) modules across VLM layers. (b) Brain Alignment Analysis: Evaluation of model-brain alignment using RSA (measuring RDM correlation) and Ridge Regression (predicting voxel-wise fMRI signals).

Representational Similarity Analysis (RSA) (Kriegeskorte et al., 2008) and Ridge Regression (Hoerl & Kennard, 1970) to quantify alignment both geometrically (via representational geometry) and predictively (via voxel-wise encoding accuracy). We conduct systematic experiments across 17 mainstream open-source VLMs, yielding three key insights:

- The CSRD framework reveals a consistent functional divergence in deep layers: Attention modules maintain strong alignment with human neural activity—acting as cognitive hubs—whereas FFN representations progressively disentangle from biological signals.
- Instruction-tuned and generative VLMs exhibit significantly higher brain alignment than contrastive baselines (e.g., CLIP), suggesting that task-driven learning better captures human-like representational structures than pure co-embedding objectives.
- Model scale alone is insufficient; architectural choices and training paradigms are dominant factors in achieving neurocognitive fidelity.

Method

Overview

To investigate the alignment between Vision-Language Models (VLMs) and human visual-cortical responses, we propose a systematic analysis framework named Component-Specific Residual Disentanglement (CSRD). This framework consists of two main stages: component-disentangled feature extraction and alignment evaluation. In the first stage, we disentangle the Attention mechanisms and Feed-Forward Networks (FFNs) of Transformer-based multimodal models to extract module-specific representations. By retaining the residual connections for Attention and stripping them for FFNs, we isolate their respective roles in information routing and nonlinear knowledge transformation. In the second stage, we evaluate the correspondence between these extracted model features and human brain activity from two complementary perspectives: Representational Similarity Analysis (RSA) to measure the similarity of representational geometries, and a Ridge Regression Encoding Model to quantify the predictive capability of model features for voxel-level brain activity. This two-stage pipeline allows us to pinpoint the specific components and mechanisms driving brain-machine alignment.

Hierarchical Disentangling and Extraction of VLM Representations

Building upon the overall framework, we detail the specific extraction process. We input pairs of visual stimuli images and a standardized prompt ("Please observe this image.") into the VLM. Let $H_l \in \mathbb{R}^{L \times D}$ be the input hidden state of the l-th Transformer block, where L is the sequence length and D is the hidden layer dimension.

Global Semantic Representations To evaluate the alignment between the image semantics output by the model and brain activity, we focus on the last layer's output of the Transformer backbone. Given the causal attention mechanism in contemporary autoregressive VLMs, the representation of the Last Token inherently encapsulates the accumulated semantic context of the entire input sequence. Specifically, we define the extraction strategy as follows:

Last Token: Directly extracts the final token (usually EOS or a special visual token) of the sequence. Due to the causal attention mechanism, this token serves as a highly abstract summary of the entire visual input:

$$\mathbf{v}_{last} = H_l[L, :] \tag{1}$$

Component-Specific Hierarchical Evolution To investigate how component-defined model representations relate to human visual-cortical representations, we analyze the internal

information flow within the Attention and FFN modules by tracking the evolution of Last Token representations.

We posit the Attention module injects global context into the current representation. Therefore, we extract the output containing the residual connection as a snapshot of information flow the integration:

$$\mathbf{F}_{attn}^{(l)} = \text{MHSA}(\text{LN}(H_l)) + H_l \tag{2}$$

For the Feed-Forward Network (FFN) module, we focus on its query and retrieval characteristics as a key-value memory network. To exclude the interference of contextual information already present in the input state, we specifically strip the residual connection and directly extract the transformation increment of the FFN, $\mathbf{F}_{ffn}^{(l)}$. This represents the new knowledge the layer attempts to "write" into the data stream, rather than the final mixed state. To isolate its specific non-linear transformation contribution (i.e., knowledge retrieval and processing), we capture its pure function output before it is added back to the residual stream:

$$\mathbf{F}_{ffn}^{(l)} = \text{FFN}(\text{LN}(\mathbf{F}_{attn}^{(l)})) \tag{3}$$

where LN denotes Layer Normalization, and $\mathbf{F}_{attn}^{(l)}$ serves as both the output of the Attention stage and the input to the FFN stage. This differentiated processing generates two independent feature matrices for each layer across N stimulus images: the Attention feature matrix $\mathbf{X}_{attn} \in \mathbb{R}^{N\times D}$ and the FFN feature matrix $\mathbf{X}_{ffn} \in \mathbb{R}^{N\times D}$. They represent the context integration Flow and Independent Functional Transformation, respectively, facilitating a fine-grained analysis of the alignment between model components and brain signals.

Mathematically, this residual stripping operation treats the FFN not as a feature refiner, but as an independent function approximator. This allows us to strictly evaluate the correlation between the 'pure' non-linear transformation learned by the network and the biological signals, without the confounding factor of the identity path.

By mathematically isolating the residual stream, we effectively decouple the 'routing' logic of Attention from the 'processing' logic of FFNs. This allows us to compare whether these component-defined representations show different correspondence patterns with human visual-cortical activity.

Representational Similarity Analysis

To evaluate whether VLMs and the human brain use similar geometric structures to encode visual information, we performed Representational Similarity Analysis (RSA).

For N stimulus images, let $\mathbf{V} \in \mathbb{R}^{N\times D}$ denote the response matrix (either fMRI voxel activities or model features). We construct the Representational Dissimilarity Matrix (RDM) $\mathbf{R} \in \mathbb{R}^{N\times N}$, utilizing the correlation distance metric. The element d_{ij} in the RDM represents the dissimilarity between the responses to the i-th and j-th images:

$$\mathbf{R}_{ij} = 1 - \frac{\text{cov}(\mathbf{v}_i, \mathbf{v}_j)}{\sigma(\mathbf{v}_i)\sigma(\mathbf{v}_j)} \tag{4}$$

where cov denotes covariance and σ denotes standard deviation (equivalent to 1 – Pearson's ρ). This yields two symmetric matrices: RDM_{brain} and RDM_{model}. The final alignment score is quantified by calculating the Spearman rank correlation between the upper triangular parts of the brain RDM and the model RDM, thereby measuring the second-order similarity between the two systems.

Ridge Regression Encoding Model

To provide a comprehensive analysis beyond the geometric similarity, and following standard protocols in neural encoding (Yamins et al., 2014), we employed a Ridge Regression Encoding Model to verify whether model features can linearly predict voxel-level brain activity.

We treat the encoding process as a linear mapping problem. Let $\mathbf{X} \in \mathbb{R}^{N\times D}$ be the disentangled model features (as regressors) and $\mathbf{Y} \in \mathbb{R}^{N\times V}$ be the fMRI responses of V voxels. Our goal is to learn a weight matrix $\mathbf{W} \in \mathbb{R}^{D\times V}$ to minimize the reconstruction error under an L_2 regularization constraint:

$$\hat{\mathbf{W}} = \min_{\mathbf{W}} \|\mathbf{Y}_{train} - \mathbf{X}_{train}\mathbf{W}\|_F^2 + \alpha\|\mathbf{W}\|_F^2 \tag{5}$$

where α is a regularization parameter optimized via cross-validation. The closed-form solution to this problem is $\hat{\mathbf{W}} = (\mathbf{X}^T\mathbf{X} + \alpha\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}$.

After training, we predict the brain responses on the held-out test set: $\hat{\mathbf{Y}}_{test} = \mathbf{X}_{test}\hat{\mathbf{W}}$. Encoding performance is evaluated using Pearson correlation coefficient or the coefficient of determination (R^2) between the predicted response $\hat{\mathbf{y}}_v$ and the ground truth response $\mathbf{y}_v$:

$$\text{Score} = \text{Corr}(\mathbf{y}_v, \hat{\mathbf{y}}_v) \quad \text{or} \quad R^2 = 1 - \frac{\sum(y_v - \hat{y}_v)^2}{\sum(y_v - \bar{y}_v)^2} \tag{6}$$

Experiments

To comprehensively evaluate the alignment between multimodal large models and human visual-semantic cognition, we conducted systematic experiments on a large-scale neuroimaging dataset, covering 17 models with different architectures, scales, and training stages.

Cognitive Vision Processing Data

We choose the Natural Scenes Dataset (NSD) as the benchmark for human cortical activity. NSD provides high-resolution whole-brain fMRI scans (high SNR) from subjects viewing 73,000 natural scene images sourced from MS COCO (Lin et al., 2015). In the present study, however, our analyses focus on the NSD-general ROI within the visual pathway. Here, ROI (region of interest) refers to a predefined subset of voxels selected to target a specific functional question. Accordingly, our conclusions concern alignment with selected visual-cortical responses rather than whole-brain or full multimodal cognition.

To ensure maximum data integrity for high-precision voxel-wise encoding, we select data from Subjects 1, 2, 5, and 7, who completed nearly all 30, 000 stimulus presentations.We

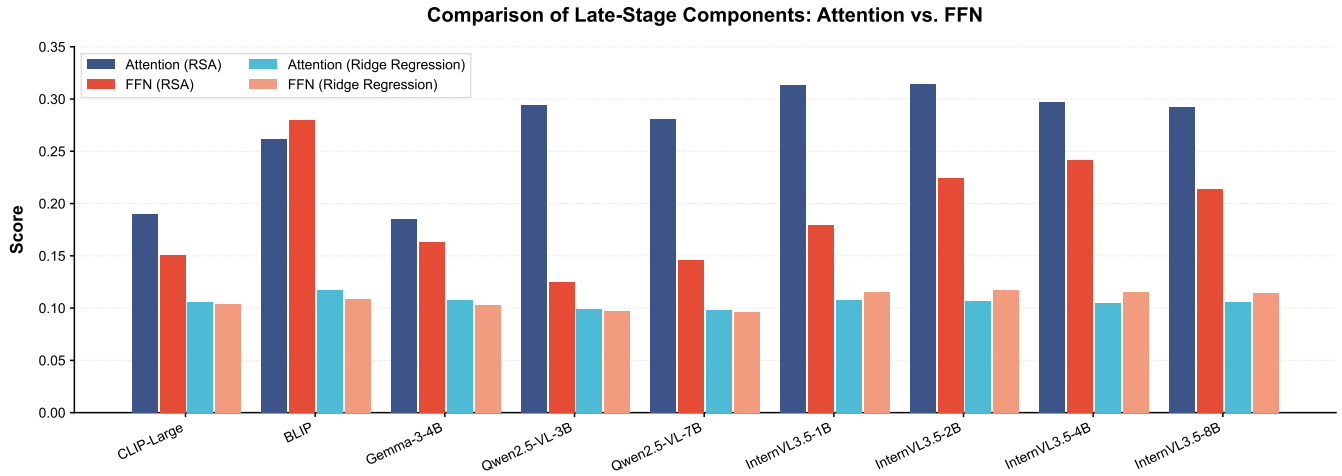

Figure 2: **Late-stage functional divergence in VLMs.** Attention mechanisms maintain high neuro-cognitive fidelity, whereas FFNs disentangle from biological signals to prioritize task-specific transformations.

focus on the NSD-general ROI (region of interest), a functionally defined mask comprising approximately 15,000 visually responsive voxels per subject. This mask covers core components of the ventral visual pathway, including early visual cortex (V1–V3) and ventral temporal cortex (VTC). Accordingly, our analyses speak most directly to alignment with selected visual-cortical responses, rather than to whole-brain or full multimodal cognition.

For preprocessing, neural responses are Z-score normalized across subjects and sessions to eliminate dimensional variances. Additionally, we filter voxels based on an SNR threshold, retaining only those with stable responses to visual stimuli for subsequent analysis.

Vision-Language Models

To investigate the impact of model scale and training paradigms on brain-machine alignment, we selected 17 representative multimodal models for evaluation.

Specifically, we chose **Qwen2.5-VL-Instruct** (3B and 7B) (S. Bai et al., 2025) and the **InternVL3.5** series (1B, 2B, 4B, and 8B) (Wang et al., 2025) as representative generative multimodal models. The InternLM series is further subdivided into three versions to dissect the contribution of different training stages: (1) Pretrain (large-scale unsupervised pre-training); (2) Instruct (instruction fine-tuning); and (3) Complete (final version with CascadeRL alignment). These models allow for a descriptive comparison of model scale and training stage with respect to brain alignment.

As a comparison, we included models based on contrastive learning paradigms, including the classic dual-tower models **CLIP-ViT-Large** and **BLIP** (Li et al., 2022), as well as the latest lightweight frontier model **Gemma-3-4B-it** (Team et al., 2024). For each model, features were extracted using the CSRD method and Z-score normalized to ensure comparability with neural response data. To facilitate a unified analysis of hierarchical representations across models with varying depths, we categorise the extracted layers into three distinct processing stages: Early, Middle, and Late. Specifically, for a model with L layers, the layers are uniformly partitioned into three equal segments based on their depth indices.

Experimental Setup and Evaluation Metrics

We evaluate brain-model alignment using two complementary approaches: Representational Similarity Analysis (RSA) and Ridge Regression encoding.

For RSA, we construct Representational Dissimilarity Matrices (RDMs) using $1 -$ Pearson correlation distance for both brain and model representations. The alignment is quantified by Spearman's rank correlation between the upper triangular parts of these RDMs. To handle large-scale matrices, we employ block-based computation and random subsampling.

For the Encoding Model, we utilize Ridge Regression with L_2 regularization to map model features to voxel-wise fMRI signals. The dataset is chronologically split into training (80%) and testing (20%) sets. The regularization hyperparameter α is optimized via 5-fold cross-validation within a logarithmic space. All experiments are fixed with Seed = 42 for reproducibility.

Alignment performance is reported on the independent test set using Pearson correlation (r) and the coefficient of determination (R^2) to measure prediction accuracy and explained variance, respectively.

Results

In this section, we evaluate brain alignment between multimodal models and the human visual cortex from four angles: Attention versus FFN behavior, layer-wise trends across network stages, effects of training stage, and a comparison between generative and contrastive models.

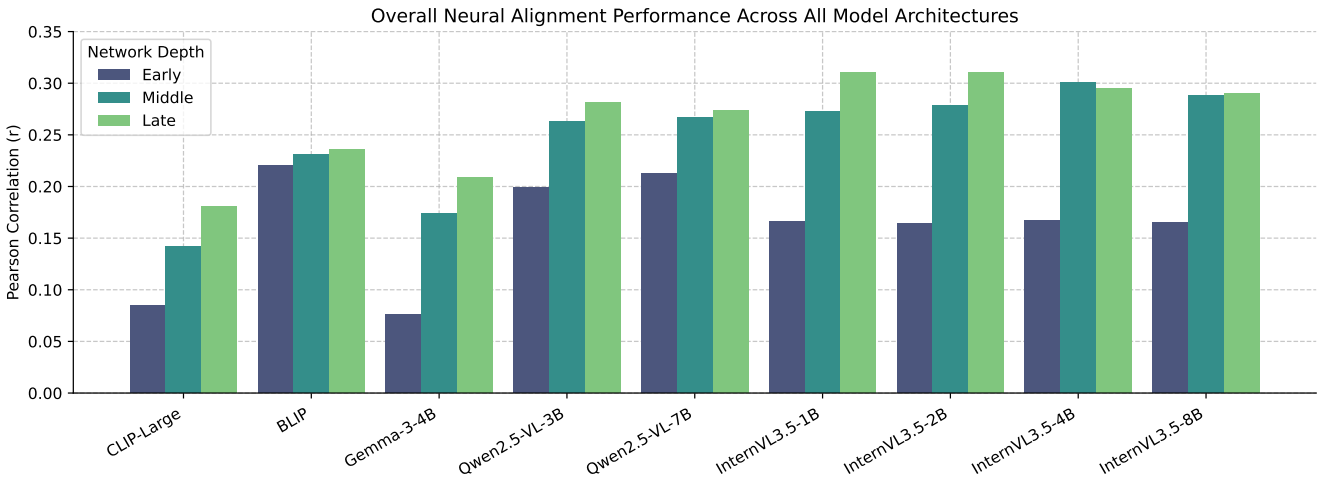

Figure 3: **Layer-wise brain alignment across VLMs.** A consistent hierarchical evolution is observed across architectures, where alignment with human fMRI signals (r) significantly improves as network depth increases from Early to Late layers.

Functional Disentangling: Divergence of Attention and FFN

To examine the mechanisms underlying brain-model alignment, we employed the CSRD framework to separate context integration (Attention) from independent transformation (FFN) pathways. Figure 2 provides a comparative view of these two components in late processing stages.

Our results show a clear component-level divergence. Attention-related representations maintain higher geometric similarity to the visual cortex in deep layers. As shown in Figure 2, Attention in Qwen2.5-VL-7B reaches an RSA correlation of $r \approx 0.28$, whereas FFN reaches $r \approx 0.15$. This pattern indicates that contextual routing in Attention remains a major contributor to neuro-cognitive alignment.

Crucially, Ridge Regression provides a complementary perspective on this divergence. In contrast to the marked RSA gap, the linear encoding performance of FFN remains comparable to Attention (both $R^2 \approx 0.10$). This suggests that deep-layer FFNs may not simply "lose" neural information; instead, they appear to reorganize representational geometry from a biologically aligned space toward a task-oriented manifold useful for generation. Accordingly, Attention better preserves the brain-like representational "shape," while FFNs contribute non-linear transformations that are functionally useful but less geometrically aligned with cortical patterns.

This pattern is broadly consistent with theories of cortical organization: Attention may serve as a shared pathway for integrating multimodal information in a biologically consistent format, whereas FFNs may implement more specialized transformations required for text generation, which can depart from purely visual biological constraints.

Hierarchical Evolution and Overall Alignment Trends

We begin by examining the layer-wise similarity between global semantic representations in VLMs and voxel-level brain activity. As shown in Figure 3, a consistent hierarchical ascent pattern emerges across different model architectures.

For most models, alignment performance increases with network depth. Taking InternVL3.5-1B as an example, the Pearson correlation of RSA (r) rises from 0.166 in the Early stage to 0.310 in the Late stage. This pattern suggests that VLMs move beyond low-level visual processing and, through deeper non-linear transformations, progressively form higher-level semantic representations that more closely resemble those in the human high-level visual cortex.

The data further shows that scaling from 1B to 2B brings visible improvements, especially in the Middle layers (from 0.273 to 0.278). However, beyond 2B parameters, alignment gains plateau or slightly decline. The Late-stage alignment of InternVL3.5-8B ($r = 0.290$) is lower than that of the smaller 1B and 2B variants. This suggests that medium-scale models may already capture much of the visual-semantic structure reflected in our brain-alignment metrics, whereas further scaling can shift representations toward less biologically aligned statistical patterns.

Impact of Training Paradigms

We examine how training stages reshape cognitive alignment by comparing the Pretrained and Instruct variants of the InternVL3.5 series (Figure 4).

The results indicate that instruction tuning consistently improves alignment in deep layers. For InternVL3.5-1B, the Instruct version ($r = 0.309$) outperforms the Pretrained version ($r = 0.274$). This pattern suggests that supervision aligned with human intent (SFT) may shift high-level representations from diffuse feature accumulation toward more

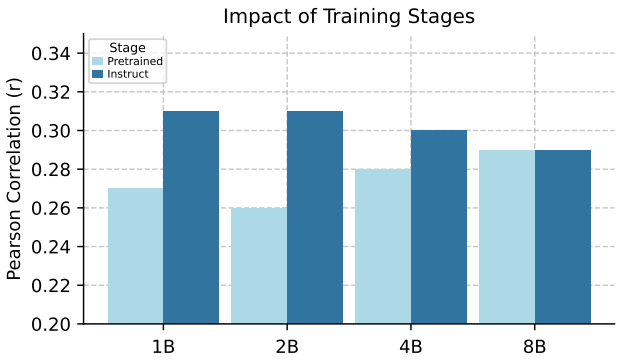

Figure 4: Instruction-tuned models consistently exhibit higher brain alignment than their pretrained counterparts across all sizes, highlighting the role of human feedback in enhancing cognitive consistency.

Table 1: **Generative vs. Contrastive Baselines.** Despite comparable R^2, RSA reveals a sharp contrast: generative models exhibit brain-like functional divergence, whereas contrastive models remain functionally homogeneous.

Model	Attention		FFN	
	RSA (r)	Ridge (R^2)	RSA (r)	Ridge (R^2)
CLIP-ViT-Large	0.190	0.106	0.151	0.104
BLIP-Base	0.262	0.117	0.280	0.109
Qwen2.5-VL-7B	0.281	0.098	0.146	0.096
InternVL3.5-1B	0.313	0.108	0.179	0.115

structured task-relevant organization, yielding stronger correspondence with human neural coding patterns.

We further hypothesize that instruction tuning, by reducing hallucinations and encouraging logical consistency, may attenuate non-biological correlations formed during large-scale pre-training. In this view, the 'alignment tax' in generative tasks can function as a regularizer, sharpening representational geometry toward patterns more consistent with efficient coding in human perception.

Furthermore, we posit that instruction tuning may do more than refine existing features; it may partially reorganize the model's latent space. During pre-training, the objective maximizes the likelihood of co-occurring tokens, which can encode redundant or noisy correlations, some of which may be less biologically plausible.

Retrospective Analysis: Generative vs. Contrastive Baselines

To contextualize the neuro-cognitive fidelity of generative VLMs, we revisited classic dual-tower contrastive baselines (CLIP and BLIP). As shown in Table 1, while Ridge Regression indicates comparable linear decodability of biological information across paradigms (e.g., Attention $R^2 \approx 0.10$ for both CLIP and InternVL3.5), RSA reveals a clear difference in component organization. Contrastive models exhibit "Functional Homogeneity," with limited internal specialization: BLIP-Base shows similarly high alignment for FFN ($r = 0.280$) and Attention ($r = 0.262$). By contrast, generative models show the "Functional Divergence" observed earlier, where Attention ($r = 0.313$) outperforms FFN ($r = 0.179$) in geometric similarity to brain responses.

This dissociation suggests that training objectives may shape representational topology. Despite comparable semantic decodability ($R^2 \approx 0.11$ across paradigms), contrastive optimization tends to preserve functional homogeneity, with BLIP's FFN ($r = 0.280$) closely tracking its Attention ($r = 0.262$). In comparison, the generative objective is associated with a clearer division of labor: InternVL3.5's Attention behaves as a high-fidelity integration pathway ($r = 0.313$), whereas FFNs show lower geometric alignment ($r = 0.179$), consistent with stronger task-oriented specialization in nonlinear transformation.

This pattern can be interpreted through a Predictive Coding perspective. Contrastive models such as CLIP are trained with global instance discrimination, which encourages components to encode high-level semantics more uniformly, matching the observed functional homogeneity. Generative models, in contrast, are optimized for next-token prediction, which may require more explicit hierarchical specialization: Attention layers maintain a global "mental sketch" of the scene (higher brain alignment), while FFNs carry out transformations needed for token prediction (task-specific deviation). Under this view, the "alignment tax" observed in deep FFNs may reflect functional specialization rather than simple degradation, analogous (though not equivalent) to the dissociation between sensory processing and symbolic manipulation in human cognition.

Conclusion

We introduce the Component-Specific Residual Disentanglement (CSRD) framework as a systematic method for comparing component-level representations in Vision-Language Models (VLMs) with human visual-cortical responses. Under the CSRD extraction scheme, we observe a consistent late-layer contrast: attention-associated representations remain more closely aligned with the human visual cortex, whereas FFN-associated representations show lower geometric alignment. We also find that, within our model set, instruction-tuned generative models tend to achieve higher alignment than the contrastive baselines examined here. Taken together, these results indicate that both model component and training stage are associated with differences in representational similarity to human brain activity. Although our findings are primarily descriptive and do not establish mechanism or causality, they provide a useful component-level perspective on VLM-brain alignment, underscore the value of finer-grained analysis beyond standard layer-wise comparisons, and motivate testable hypotheses for future studies of model-brain correspondence in multimodal systems.

References

Alain, G., & Bengio, Y. (2016). Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

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., et al. (2022). A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, *25*(1), 116–126.

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al. (2023). Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., ... Lin, J. (2025). Qwen2.5-vl technical report. https://arxiv.org/abs/2502.13923

Belinkov, Y. (2021). Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, *48*, 207–219. https://api.semanticscholar.org/CorpusID:236924832

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. https://arxiv.org/abs/2303.12712

Caucheteux, C., Gramfort, A., & King, J.-R. (2022). Deep language algorithms predict semantic comprehension from brain activity. *Scientific Reports*, *12*. https://doi.org/10.1038/s41598-022-20460-9

Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al. (2024). Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 24185–24198.

Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). What does bert look at? an analysis of bert's attention. https://arxiv.org/abs/1906.04341

Geva, M., Schuster, R., Berant, J., & Levy, O. (2021). Transformer feed-forward layers are key-value memories. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 5484–5495.

Guclu, U., & van Gerven, M. A. J. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005–10014. https://doi.org/10.1523/jneurosci.5023-14.2015

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. https://arxiv.org/abs/2001.08361

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4. https://doi.org/10.3389/neuro.06.004.2008

Li, J., Li, D., Xiong, C., & Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. https://arxiv.org/abs/2201.12086

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., & Dollár, P. (2015). Microsoft coco: Common objects in context. https://arxiv.org/abs/1405.0312

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. *Advances in neural information processing systems*, *36*, 34892–34916.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. *International conference on machine learning*, 8748–8763.

Schrimpf, M., Blank, I., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *bioRxiv*. https://doi.org/10.1101/2020.06.26.174482

Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., Tafti, P., Hussenot, L., Sessa, P. G., Chowdhery, A., Roberts, A., Barua, A., Botev, A., Castro-Ros, A., Slone, A., ... Kenealy, K. (2024). Gemma: Open models based on gemini research and technology. https://arxiv.org/abs/2403.08295

Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., Wang, Z., Chen, Z., Zhang, H., Yang, G., Wang, H., Wei, Q., Yin, J., Li, W., Cui, E., ... Luo, G. (2025). Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. https://arxiv.org/abs/2508.18265

Wu, T., Yang, G., Li, Z., Zhang, K., Liu, Z., Guibas, L., Lin, D., & Wetzstein, G. (2024). Gpt-4v (ision) is a human-aligned evaluator for text-to-3d generation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22227–22238.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, *111*(23), 8619–8624.