CROSS-MODAL NEURAL ENCODING IS ASYMMETRIC: EVIDENCE FROM VISION-LANGUAGE MODELS AND PAIRED TEXT-IMAGE FUNCTIONAL MRI

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) have been shown to predict neural responses to visual stimuli better than unimodal models, which has been interpreted as evidence that VLMs and the human brain converge on shared, cross-modal representations. However, existing studies supporting this claim have critical limitations due to the functional magnetic resonance imaging (fMRI) datasets they use, because they often present a single modality, bind modalities inseparably (e.g., a video with simultaneous audio), present mismatched stimuli across modalities, or restrict paired stimuli to isolated concepts. We address these limitations by using a recent phrase- and scene-level text-image fMRI dataset built from paired stimuli: each stimulus exists as both a caption and an image, but any given participant encounters it in only one form. Therefore, cross-modal prediction cannot be attributed to simultaneous presentation, enabling a direct test of whether VLMs and the brain share cross-modal representations. Across multiple VLMs and unimodal baselines, we find a consistent asymmetry: text embeddings reliably predict image-evoked brain activity, whereas image embeddings are largely unable to predict text-evoked brain activity above chance level. We then use a residualization analysis to show that removing the cross-modally shared information from VLM embeddings selectively reduces cross-modal encoding performance to a larger extent than within-modality encoding performance. Mapping the embedding spaces onto each other directly, we find this same asymmetry: text embeddings predict image embeddings more reliably than vice versa. Taken together, these results indicate that VLMs and the brain do share cross-modal information, but not symmetrically: cross-modal information may be more directly accessible from text embeddings than from image embeddings.

1 INTRODUCTION

Neuroscientists increasingly use the internal representations of computational models as hypotheses about how the brain encodes information. These hypotheses can be tested via neural encoding and decoding models (Mitchell et al., 2008; Naselaris et al., 2011; Huth et al., 2016; Oota et al., 2022), i.e., predicting brain activity from stimuli and vice versa, often by means of functional magnetic resonance imaging (fMRI). Past work has thoroughly characterized visual representations: successive layers of deep vision models predict successive stages of the visual system, from low-level features in early visual cortex to semantic categories in high-level visual cortex (Yamins et al., 2014; Güçlü & Gerven, 2015; Wen et al., 2018). In language, there is a similar correspondence for semantic representations, where successive layers of language models are able to predict neural representations in the language network, spanning temporal, parietal and prefrontal regions (Toneva & Wehbe, 2019; Schrimpf et al., 2021; Caucheteux & King, 2022). Yet, in both cases, single-modality language or vision models are compared to brain activity evoked by that same modality. Semantic cognition, in contrast, is widely thought to rely on modality-invariant representations (Damasio, 1989; Lambon Ralph et al., 2017). Vision-language models (VLMs) are inspired by that idea, and a correspondence across modalities is the explicit target of joint vision-language training (Li et al., 2025). However, the degree to which cross-modal (i.e., shared across input modalities) representations in the brain correspond to cross-modal representations in VLMs remains an open question.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Multimodal neural encoding studies have begun to examine this question, motivated by the finding that VLMs outperform unimodal baselines in neural encoding (Popham et al., 2021; Tang et al., 2023; Oota et al., 2025; Bavaresco et al., 2026b), but they share a fundamental limitation: existing fMRI datasets do not enable a controlled and cleanly separable cross-modal evaluation. Specifically, most datasets either (i) do not have paired stimuli across modalities (Huth et al., 2016; Chang et al., 2019; Allen et al., 2022), (ii) bind modalities together inseparably (e.g., video and audio) (Hanke et al., 2014; Oota et al., 2025), (iii) present mismatched stimuli across modalities (Popham et al., 2021; Tang et al., 2023), or (iv) restrict paired stimuli to isolated concepts such as words and images of single objects rather than full phrases or naturalistic scenes (Mitchell et al., 2008; Pereira et al., 2018). Consequently, the majority of the existing research relies on indirect modality-mismatched proxies, such as large language model (LLM)-generated captions for visual stimuli (Doerig et al., 2025), VLM-generated images for textual stimuli (Saha et al., 2025), or comparing VLM representations against single-modality fMRI data (Oota et al., 2022). Hence, there is a risk of conflating two distinct settings: multimodal neural encoding, where a fused representation of multiple modalities is compared to brain activity, and cross-modal neural encoding, where a representation derived from one modality must predict brain activity elicited by another. As a result, text and image embeddings in VLMs have not been directly compared to modality-matched brain activity of the same content.

In this work, we address this gap by extending VLM-based neural encoding beyond single words and isolated objects, using a recent fMRI dataset in which each participant viewed half the stimuli as scene-level images and half as phrase-level captions. This pairing enables us to investigate three research questions (RQs): **(RQ1) Do VLM representations share information with the brain across modalities, specifically in cross-modal neural encoding settings?** We adopt a 2×2 design that crosses the embedding modality with the fMRI modality, evaluating within-modality encoding, i.e., embeddings predicting $\rightarrow$ fMRI of the same modality (text embeddings $\rightarrow$ text-evoked fMRI; image embeddings $\rightarrow$ image-evoked fMRI) alongside cross-modal encoding (text embeddings $\rightarrow$ image-evoked fMRI; image embeddings $\rightarrow$ text-evoked fMRI) to identify which embeddings, if any, carry modality-invariant information shared with the brain. **(RQ2) Is the encoding performance between VLM embeddings and the brain attributable to cross-modal information?** Extending a residual-encoding approach (Oota et al., 2023; 2025), we remove from each embedding the component linearly predictable from the other modality and re-evaluate encoding, testing whether cross-modal information carries encoding performance. **(RQ3) Can VLM-based text and image embeddings predict one another equally well?** We directly predict the two types of VLM embeddings from one another (text $\Rightarrow$ image and image $\Rightarrow$ text), using both linear and nonlinear regression settings, to quantify how much cross-modal information is accessible via each mapping and to examine whether cross-modal predictability is directional within VLMs themselves. Our main contributions can be summarized as follows:

1. **Cross-modal neural encoding of VLMs at the level of natural scenes and phrases.** Leveraging a recent paired text-image fMRI dataset, we move beyond the single-concept granularity of existing studies by using phrase- and scene-level stimuli to perform a cleanly separated cross-modal evaluation of VLM embeddings against human brain activation to text and images (§3.3).
2. **Evidence of asymmetric cross-modal neural encoding across VLMs.** Across multiple VLMs, we show that, in the same participants, text embeddings reliably predict image-evoked brain activity, but image embeddings largely do not predict text-evoked brain activity above chance level, an asymmetry consistent across subjects and model families (§4.1).
3. **Attribution of the cross-modal discrepancy to modality-shared information.** We show that removing the shared component from image and text embeddings reduces cross-modal encoding more than within-modality encoding. This identifies modality-shared information as the principal carrier of cross-modal neural encoding performances (§4.2). Mapping the embedding spaces onto each other directly, we find that text embeddings predict image embeddings more reliably than vice versa, mirroring the directionality of the neural encoding results (§4.3).

2 Related Work

A growing body of work has examined cross-modal and multimodal neural encoding in VLMs. For instance, Doerig et al. (2025) leverage embeddings of LLM-generated captions to predict brain activity evoked by visual stimuli, showing that the integration of complex higher-level information in

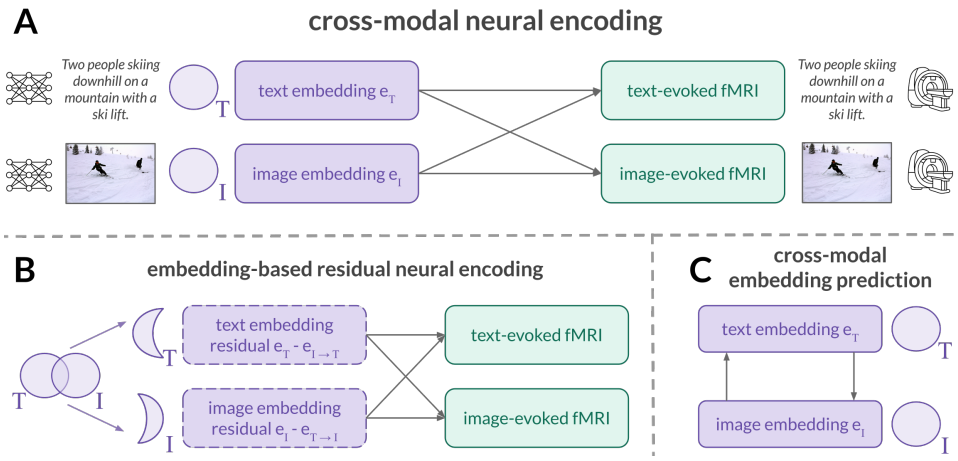


Figure 1: **Methods overview. (A) Cross-modal neural encoding**: we perform within-modality encoding and cross-modal encoding to examine how vision-language model (VLMs) embeddings correspond to brain representations of each modality (RQ1). **(B) Neural encoding with residualized embeddings**: the contents of text (T) and image (I) embeddings are depicted as overlapping sets, whose intersection is the cross-modal component. We use ridge regression (e.g., $e_{I\to T}$) to identify and subtract the cross-modal component (e.g., $e_T - e_{I\to T}$) yielding residualized embeddings, i.e., the isolated crescents: T, I, (see Table 1). These serve as encoding input for text- and image-evoked fMRI (RQ2). See Appendix A.7 for the analogous residualization applied to the fMRI. **(C) Cross-modal embedding-only prediction**: we predict one embedding modality from the other in both directions, quantifying directional cross-modal predictability within VLMs alone (RQ3).

text embeddings is driving the correspondence to image-evoked brain activity. Artificially generated captions further surpass human-generated ones in predicting image-evoked brain activity (Bavaresco et al., 2026a). Similarly, Popham et al. (2021) compare semantic maps elicited by silent movies and different narrative stories, indicating a converging boundary of linguistic and visual representations in the brain, and Saha et al. (2025) demonstrate that averaged image embeddings across multiple artificially generated images align with neural representations from the human language cortex. Oota et al. (2022) and Bavaresco et al. (2026b) demonstrate that VLMs outperform unimodal models in neural encoding settings based on single-modality fMRI, and Tang et al. (2023) demonstrate transfer across fMRI responses to stories and movies. Closest to our proposed approach, Oota et al. (2025) residualize cross-modal against unimodal representations, but because both modalities are presented jointly, their shared representation cannot be separate from the fact that they co-occurred. Finally, studies employing paired text-image stimuli have, to date, been restricted to the level of single words and object images (Mitchell et al., 2008; Pereira et al., 2018). We complement these approaches by introducing a strictly parallel cross-modal setting at the level of matched phrases and naturalistic scenes. Beyond using caption embeddings to predict image-evoked fMRI responses, this setup allows us to test how image embeddings of naturalistic scenes (rather than artificially generated images) encode text-evoked fMRI responses of matching stimuli, a direction not investigated to date. Moreover, cognitive neuroscience research offers various accounts on the integration of modalities, ranging from distributed convergence zones (Damasio, 1989) and a broadly distributed modality-agnostic semantic network (Binder et al., 2009) to a hub in the anterior temporal lobe (Lambon Ralph et al., 2017), though grounded-cognition views hold that meaning remains tied to modality-specific systems (Barsalou, 2008). By leveraging parallel text-image fMRI data, we investigate whether the information shared across the text and vision modalities is what aligns VLMs with brain activity.

3 Methods

3.1 Data

Our analyses are based on two datasets: a paired text-image fMRI dataset for the neural encoding analyses (RQ1, RQ2), and a larger text-image dataset for the cross-modal embedding prediction

(RQ3). For neural encoding, we use a paired text-image fMRI dataset (Balabin et al., 2026), in which eight participants viewed 252 text-image pairs drawn from the Common Objects in Context-Actions (COCO-A) dataset (Ronchi & Perona, 2015). Each pair was shown as a caption to half of the participants and as the corresponding image to the other half, ensuring no within-subject pairing between a caption and its matching image. Each stimulus was repeated six times across three scanning sessions. Preprocessing was performed with fMRIPrep (Esteban et al., 2019), and single-trial blood-oxygenation-level–dependent (BOLD) response amplitudes (i.e., type-D betas) were estimated using general linear models (GLMs) in GLMsingle (Prince et al., 2021), and z-scored per voxel within each run prior to encoding. All analyses were conducted in subject-native space to preserve individual voxel geometry, and a whole-brain mask was applied throughout. Further details on the image acquisition, stimulus selection and the experimental procedure are provided in Balabin et al. (2026). For the cross-modal embedding prediction (§3.5), which does not involve fMRI, we use a subset of the overlap of the Visual Genome (VG) (Krishna et al., 2017) and Common Objects in Context (COCO) (Lin et al., 2014) datasets, comprising $46,944$ image-caption pairs. This extends the embedding-prediction analysis beyond the 252 fMRI stimuli, yielding a substantially larger and more reliable estimate of cross-modal predictability. In all analyses, weights are fit using five-fold cross-validation over stimuli.

3.2 Models

We evaluate a range of vision-language (VL) and unimodal models spanning various architectures, training objectives, and parameter counts to assess the generalizability of our findings. VL models include Qwen3.5-VL at 0.8B, 2B, 4B, and 9B parameters (Bai et al., 2025), InternVL3.5 at 1B, 2B, 4B, and 8B parameters (Wang et al., 2025), and the open contrastive language-image pre-training (OpenCLIP) model in its ViT-L/14 (L) and ViT-H/14 (H) variants (Ilharco et al., 2021). To dissociate effects of joint VL training from those of unimodal pre-training, we additionally evaluate unimodal image embeddings, namely the self-distillation with no labels (DINOv2) (Oquab et al., 2024) and image-based Joint-Embedding Predictive Architecture (I-JEPA) models (Assran et al., 2023) at two model scales each, and unimodal language embeddings, namely Pythia (1.4B, 6.9B) (Biderman et al., 2023) and Open Pre-Trained Transformers (OPT) (1.3B, 2.7B) (Zhang et al., 2022). For each model, embeddings are extracted from a middle-most layer, a depth at which representations are most predictive of neural activity and yield the strongest general-purpose features (Schrimpf et al., 2021; Caucheteux & King, 2022; Bavaresco et al., 2026a), and subsequently pooled by mean aggregation over spatial patches or non-padding tokens to yield one fixed-dimensional vector per stimulus. In Appendix A.1, we additionally compare neural encoding performances across all layers.

3.3 Cross-Modal Neural Encoding (RQ1)

To address the first research question (RQ1), we leverage a 2×2 experimental design that crosses the source modality of the VLM embedding (image embeddings versus text embeddings) with the modality of the evoked fMRI response (image-evoked versus text-evoked), yielding two within-modality conditions (image embeddings→image fMRI, text embeddings→text fMRI) and two cross-modal conditions (image embeddings→text fMRI, text embeddings→image fMRI). The within-modality conditions measure the neural encoding performance when embedding and stimulus modalities match, while the cross-modal conditions quantify how much of this performance is retained when they are swapped (Figure 1A).

For each participant-condition pairing, we fit a voxel-wise linear encoding model that maps VLM embeddings to single-trial fMRI responses. All model embeddings are first reduced to 50 principal components via PCA, both to match the widely varying embedding dimensionalities across models, so that performance differences reflect representational content rather than dimensionality (Antonello et al., 2023), and because the brain-relevant information in such representations has been shown to be low-dimensional (Antonello et al., 2021). PCA is applied in all neural encoding settings (this section and § 3.4), but not the cross-modal embedding prediction (see § 3.5). We then fit fractional ridge regression (Rokem & Kay, 2020) and select its regularization hyperparameter per voxel from the range $\{0.1, 0.25, 0.5, 0.75, 1.0\}$ via nested cross-validation. The outer loop uses five-fold cross-validation on stimulus identity to obtain the encoding accuracy. At test time (but not during training), repeated-stimulus responses are averaged before evaluation. Encoding accuracy is measured by the Pearson correlation r between predicted and measured fMRI responses per voxel. We follow the

approach presented in Allen et al. (2022), described in more detail in Appendix A.2, to compute the noise ceiling. All reported accuracies are computed over the top 20% of voxels based on those with the highest noise ceiling within each fMRI modality; we assess the robustness of this selection in two replications, using the top 40% and 60% most reliable voxels (Appendix A.3) and a single voxel set shared between the two modality-specific selections (Appendix A.4). The spatial distribution of the voxels with the highest noise ceiling is shown in Appendix A.5 for each subject.

Group-level performance is summarized as the mean across subjects, and subject-wise accuracies are reported in Appendix A.6. Statistical significance is assessed via a stimulus-label permutation test: for each subject, the mapping (i.e., correspondence) between embeddings and fMRI responses is shuffled across stimuli and the full encoding is re-evaluated to build a subject-level null distribution of the mean accuracy across the selected voxels ($n_{\text{perm}} = 1{,}000$ permutations). Following the bootstrap method introduced in Stelzer et al. (2013), these subject-level null distributions are combined into a group-level null distribution by repeatedly averaging draws across subjects. Let r_{real} denote the observed across-subject mean accuracy, and r_{null} are repeated draws from this group-level null distribution; the significance of r_{real} is assessed by $p = (\#\{r_{\text{null}} \geq r_{\text{real}}\} + 1)/(n_{\text{draws}} + 1)$, with $n_{\text{draws}} = 10{,}000$. The resulting p-values are corrected for multiple comparisons across all model $\times$ condition tests using the Benjamini-Hochberg false discovery rate (BH-FDR) procedure. All pairs of within- and cross-modal encoding conditions are compared using two-sided paired Wilcoxon signed-rank tests on the per-model normalized means, with the resulting p-values likewise BH-FDR-corrected across the six condition pairs.

3.4 Residual Neural Encoding (RQ2)

To test whether cross-modal neural encoding relies specifically on cross-modally shared information in the embeddings (RQ2), we extend the residual-encoding approach presented in Oota et al. (2023; 2025) (see Figure 1B). Here, we choose a residualization approach over variance partitioning (Bonner & Epstein, 2018; Shimizu et al., 2025) because the residualized features have the same dimensionality as the features in the baseline encoding model. Variance partitioning instead requires concatenating modalities into a joint model, combining the dimensions of both feature sets, which alters the regularization and complicates comparison to a baseline model.

Intuitively, one can conceptualize the information content of the text and image embeddings as two overlapping circles in a Venn diagram (Figure 1B, top right). The overlapping intersection represents the shared information that can be linearly translated between the two modalities (cf. the linear mapping between modalities presented in Oota et al., 2025). Our residualization approach subtracts this intersection, isolating only the modality-specific information, corresponding to the isolated crescent shapes ($\mathbb{C}_{\text{T}}$, $\mathbb{D}_{\text{I}}$) in Figure 1B (left). By feeding only this modality-specific information into the encoding model, we can test whether cross-modal neural encoding relies entirely on the shared intersection or not. Specifically, if cross-modal correspondence is driven by this modality-shared information, removing the component one modality linearly predicts from the other should reduce cross-modal neural encoding performance more than within-modality performance, the latter of which can additionally draw on isolated modality-specific information.

PCA is applied before residualization to keep the representations comparable to the baseline neural encoding setting, so that the resulting drop in accuracy is attributable to the removal of modality-shared information alone. We compute the shared component directly using the cross-modal ridge regression and remove it from the VLM embedding that serves as the encoding input. The regression is fit on the PCA-reduced embeddings of the training stimuli. The analogous residualization applied to the fMRI response is reported in Appendix A.7.

Let m denote the modality of the encoding input and m' the other modality, $\mathbf{e}_m$ and $\mathbf{e}_{m'}$ the PCA-reduced embeddings of the same stimulus in these two modalities, and $\mathbf{y}$ the fMRI response predicted from $\mathbf{e}_m$. We fit a ridge regression model $\mathbf{W}$ (with regularization strength $\alpha = 1.0$) to predict $\mathbf{e}_m$ from $\mathbf{e}_{m'}$ on the training stimuli, form the prediction $\hat{\mathbf{e}}_m = \mathbf{W}\mathbf{e}_{m'}$ (analogous to Oota et al., 2025), and keep only the residual $\tilde{\mathbf{e}}_m$ it does not explain:

$$\tilde{\mathbf{e}}_m = \mathbf{e}_m - \hat{\mathbf{e}}_m, \qquad \hat{\mathbf{e}}_m = \mathbf{W}\mathbf{e}_{m'}, \qquad \mathbf{W} = \arg\min_{\mathbf{W}} \sum_{i \in \text{train}} \left\| \mathbf{e}_{m,i} - \mathbf{W}\mathbf{e}_{m',i} \right\|_2^2 + \alpha \|\mathbf{W}\|_F^2 \quad (1)$$

Condition		Prediction	Residual input $\tilde{\mathbf{e}}$		Encode → Target
text→text	(within)	$\hat{\mathbf{e}}_{\text{text}} = \mathbf{W}\mathbf{e}_{\text{image}}$	$\tilde{\mathbf{e}}_{\text{text}} = \mathbf{e}_{\text{text}} - \hat{\mathbf{e}}_{\text{text}}$	$\mathbb{C}_{\text{T}}$	$\tilde{\mathbf{e}}_{\text{text}} \rightarrow \mathbf{y}_{\text{text}}$
image→image	(within)	$\hat{\mathbf{e}}_{\text{image}} = \mathbf{W}\mathbf{e}_{\text{text}}$	$\tilde{\mathbf{e}}_{\text{image}} = \mathbf{e}_{\text{image}} - \hat{\mathbf{e}}_{\text{image}}$	$\mathbb{D}_{\text{I}}$	$\tilde{\mathbf{e}}_{\text{image}} \rightarrow \mathbf{y}_{\text{image}}$
image → text	(cross)	$\hat{\mathbf{e}}_{\text{image}} = \mathbf{W}\mathbf{e}_{\text{text}}$	$\tilde{\mathbf{e}}_{\text{image}} = \mathbf{e}_{\text{image}} - \hat{\mathbf{e}}_{\text{image}}$	$\mathbb{D}_{\text{I}}$	$\tilde{\mathbf{e}}_{\text{image}} \rightarrow \mathbf{y}_{\text{text}}$
text → image	(cross)	$\hat{\mathbf{e}}_{\text{text}} = \mathbf{W}\mathbf{e}_{\text{image}}$	$\tilde{\mathbf{e}}_{\text{text}} = \mathbf{e}_{\text{text}} - \hat{\mathbf{e}}_{\text{text}}$	$\mathbb{C}_{\text{T}}$	$\tilde{\mathbf{e}}_{\text{text}} \rightarrow \mathbf{y}_{\text{image}}$

Table 1: **Embedding-based residualization conditions.** For each of the four conditions, the modality-shared component $\hat{\mathbf{e}}$ is obtained by ridge-regressing the embedding $\mathbf{e}$ on the embedding of the respective **other** modality (both PCA-reduced), fit on the training stimuli only. The encoding input is then replaced by the residual $\tilde{\mathbf{e}} = \mathbf{e} - \hat{\mathbf{e}}$, which retains only the modality-specific part of the embedding, and used to predict the unmodified fMRI target $\mathbf{y}$. Conceptually, the residual input $\tilde{\mathbf{e}}$ corresponds to the isolated modality-specific crescents ($\mathbb{C}_{\text{T}}$, $\mathbb{D}_{\text{I}}$) illustrated in Figure 1B.

The residualization is run for all four conditions of the 2×2 design (see Table 1). The residual $\tilde{\mathbf{e}}_m = \mathbf{e}_m - \hat{\mathbf{e}}_m$ (with $\hat{\mathbf{e}}_m = \mathbf{W}\mathbf{e}_{m'}$) keeps only the modality-specific part of the embedding input and replaces $\mathbf{e}_m$ to predict $\mathbf{y}$. We report the absolute neural encoding scores based on the residualized embeddings (r/r_{NC}) as the primary metric. A larger drop in the cross-modal than in the corresponding within-modality condition would constitute evidence that modality-shared information is the principal carrier of cross-modal neural encoding in VLMs, whereas a comparable drop in both would indicate that the same component also underlies the within-modality case. Two complementary views, namely the absolute difference between the residualized and baseline noise-ceiling-normalized accuracies and the relative change, are reported in Appendix A.8.

3.5 Cross-Modal Embedding Prediction (RQ3)

To better understand the residual neural encoding setting of the previous section, we measured exactly how much modality-shared information we are subtracting from each embedding. We therefore ask whether the two embeddings are mutually predictable (RQ3), and whether this predictability works better in one direction versus the other (Figure 1, bottom right). To estimate cross-modal predictability reliably, we decouple this analysis from the fMRI stimuli and instead use the larger VG-COCO image-caption dataset (see §3.1). For each VLM, and the unimodal baselines, where applicable, we predict the pooled embedding of one modality from the pooled embedding of the other for the same stimulus, in both directions (text ⇒ image and image ⇒ text). We evaluate a linear ridge regression model, its regularization strength is selected through nested cross-validation: within each training fold, $\alpha \in \{10^{-3}, 10^{-2}, \ldots, 10^{6}\}$ is chosen separately for each direction and encoder pair by maximizing the mean Pearson correlation across an inner five-fold cross-validation. To test for potential nonlinear cross-modal relationships, Appendix A.9 replaces the linear map with a multilayer perceptron. Both predictors are evaluated under an outer five-fold cross-validation, and prediction quality is measured by the Pearson correlation between predicted and true embeddings.

4 Results

4.1 Cross-Modal Neural Encoding Results (RQ1)

Within-modality encoding is reliable for both embedding modalities (see Figure 2): image embeddings predict image-evoked brain activity most strongly ($r/r_{\text{NC}} = 0.079$), and text embeddings predict text-evoked brain activity at $r/r_{\text{NC}} = 0.032$, both significantly above chance. However, the cross-modal conditions are asymmetric: text embeddings predict image-evoked brain activity significantly above chance level ($r/r_{\text{NC}} = 0.040$, and on par with or above within-modality text encoding), whereas image embeddings largely fail to predict text-evoked brain activity ($r/r_{\text{NC}} = 0.004$; the image embeddings→text-evoked fMRI neural encoding results are not significant for most tested models based on permutation testing, except for InternVL3.5 2B, Qwen3.5-VL 9B, and OpenCLIP large). This asymmetry is consistent across all three VLM families (Qwen3.5-VL, InternVL3.5, OpenCLIP), across subjects (see Appendix A.6), and the unimodal vision and language model baselines (DINOv2, I-JEPA, Pythia, OPT). It remains consistent even for different voxel-selection thresholds (Appendix A.3), including when the analyses are conducted on a shared set of voxels across image-based and text-based fMRI responses (Appendix A.4). Notably, no single model outperforms all others in terms

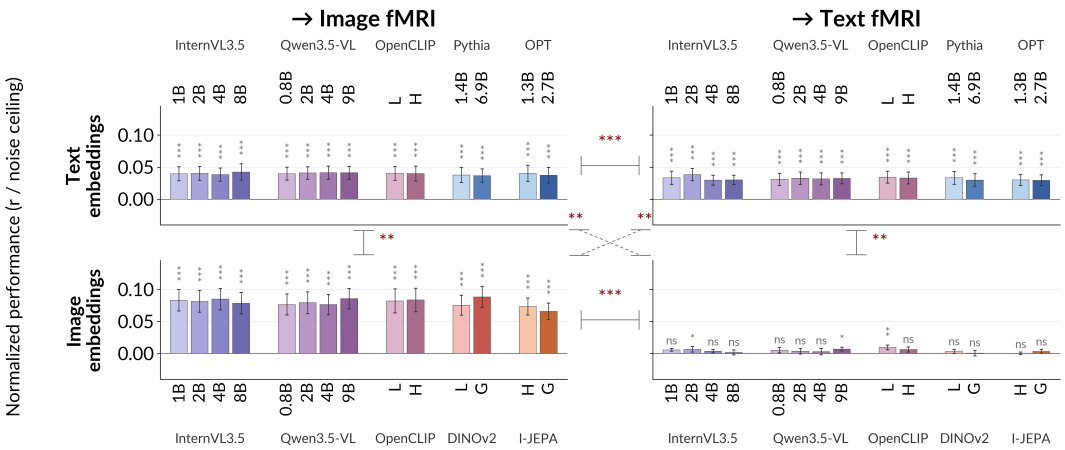


Figure 2: **Cross-modal neural encoding (RQ1).** Noise-ceiling-normalized encoding accuracy (r/r_{NC}, group mean over the eight subjects, computed over the top 20% of voxels by noise ceiling) for each model, grouped into the four conditions: image embeddings→image functional MRI (fMRI) and text embeddings→text fMRI (within-modality), as well as image embeddings→text fMRI and text embeddings→image fMRI (cross-modal). Bars are colored by model type and family; unimodal vision and language baselines are used in the two conditions matching their modality only. Significance levels for the individual models are assessed via permutation testing (***p<0.001, **p<0.01, *p<0.05, ns = not significant), and between-group brackets indicate differences between conditions (see §3.3). All p-values are corrected using the Benjamini-Hochberg False Discovery Rate (BH-FDR) method. L, H and G denote ViT-L/14, ViT-H/14 and ViT-G vision encoder backbones.

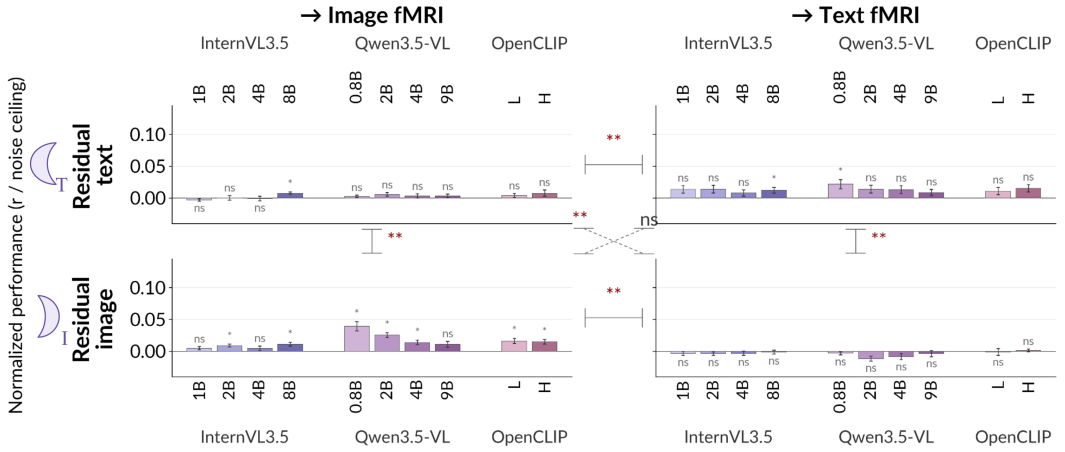


Figure 3: **Embedding-based residual neural encoding (RQ2).** Absolute noise-ceiling-normalized encoding accuracy scores after removing the component of each embedding linearly predictable from the other modality, shown per model and per condition. The residualization setups are listed in Table 1. Error bars denote across-subject variability. The corresponding relative and absolute differences between the residualized and baseline neural encoding accuracies (see Figure 2) are provided in Appendix A.8. L and H denote the OpenCLIP ViT-L/14 and ViT-H/14 variants, respectively.

of neural encoding performance. All in all, a consistent directionality emerges: **text embeddings can reliably predict both image-evoked and text-evoked brain activity, but image embeddings largely cannot predict text-evoked brain activity at an above-chance level**.

4.2 Residual Neural Encoding Results (RQ2)

We now turn to our residual analysis, where we remove from modality a the component that is linearly predictable from modality b. Using residualized embeddings reduces encoding performance in all

four conditions, but to differing degrees (see Figure 3; absolute and relative differences with respect to the baseline are reported in Appendix A.8). Removing information in image embeddings predictable from text embeddings affects the cross-modal condition significantly more than the within-modality condition (see the lower row in Figure 3). Similarly, removing the shared component from text embeddings leads to significantly larger decreases in performance for the text embeddings→image-evoked fMRI condition than for the text embeddings→text-evoked fMRI condition (see the upper row in Figure 3). Residualizing the neural responses themselves rather than the embeddings reproduces an analogous pattern (Appendix A.7), providing converging evidence that cross-modal encoding performance largely depends on modality-shared information. Although absolute within-modality performance does not differ significantly, more encoding models remain above chance in the image embeddings→image-evoked fMRI than in the text embeddings→text-evoked fMRI condition (seven versus two), suggesting that **image embeddings retain richer modality-specific information, whereas text embeddings may rely more on modality-shared information**.

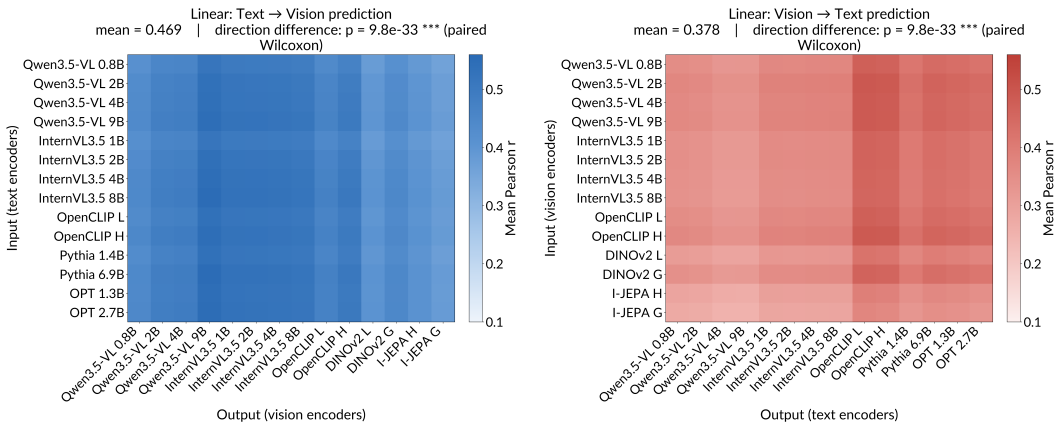
4.3 Cross-Modal Embedding Prediction Results (RQ3)

The asymmetry of the cross-modal neural encoding results reported in §4.1 is also reflected within the VLM embedding representations themselves (see Figure 4). When predicting image and text embeddings from one another, text⇒image prediction yields substantially higher correlations than image⇒text (mean Pearson $r = 0.47$ vs. 0.38, $p < 0.001$). Performances are more consistent across columns than rows, suggesting that the predictability across the two types of embeddings depends on the target embedding model (columns of the heatmaps shown in Figure 4) more than on the source embedding model used to predict it (rows of the heatmaps shown in Figure 4). The nonlinear MLP yields near-identical values (see Appendix A.9), indicating that the observed asymmetry in cross-modal prediction scores is not a mere artifact of the inherent limitations of linear projections. Thus, across the tested VLMs and unimodal baselines, **the information present in image embeddings is consistently more recoverable from text embeddings than vice versa**, mirroring the directional correspondence reported in §4.1.

5 Discussion

Overall, our findings yield a striking asymmetry: a shared and likely cross-modal semantic code is recoverable from text embeddings, which predict image-evoked brain activity as well as text-evoked brain activity, but is not recoverable from image embeddings, which cannot reliably predict text-evoked brain activity at an above-chance level. The successful encoding from text embeddings→image fMRI is in line with previous findings on other datasets (Doerig et al., 2025; Bavaresco et al., 2026b). By contrast, the absence of above-chance encoding performances for the understudied image embeddings→text fMRI direction has, to the best of our knowledge, not been reported before. Markedly, our results contrast with Saha et al. (2025), who reported that averaged embeddings from multiple artificially generated images predicted neural responses in the language cortex. However, their regression was applied to full-dimensional embeddings averaged across multiple generated images per stimulus rather than a PCA-reduced embedding of a single naturalistic scene. The difference between their findings and ours may indicate that an image embeddings→text fMRI encoding setting only succeeds if the employed image embeddings sufficiently capture the semantic gist. In Saha et al. (2025), each generated image conveyed one prominent concept, even when created from a whole sentence. Our naturalistic scenes, by contrast, are semantically richer.

Two independent analyses provide converging additional explanations for our reported result. First, the observed asymmetry is unlikely to be an artifact of the brain measurement alone, as it would also impact the mutual predictability of the two embedding modalities. This asymmetry in the VLM embeddings themselves may be related to information imbalance (Schrodi et al., 2024). For example, explicit symbolic structure in language may facilitate representing higher-level cross-modal semantics in the language rather than the vision part of a VLM. While an analysis of visual tokens in vision transformers revealed some similarities to natural languages, it also highlighted important differences, in particular with regard to the absence of a hierarchical organization and indications for dominating intermediate rather than higher-order levels of granularity (Chan et al., 2024). This absence of higher-order (semantic) information may explain why image embeddings predict text embeddings less robustly than vice versa and why they do not reliably predict text-evoked fMRI above



(a) Text-to-image: predicting the pooled vision embedding of the target model (columns) from the pooled text embedding of the source model (rows).

(b) Image-to-text: predicting the pooled text embedding of the target model (columns) from the pooled vision embedding of the source model (rows).

Figure 4: **Cross-modal embedding prediction (RQ3).** Pearson r of linearly predicting the pooled embedding of one modality from the pooled embedding of the other modality for the same stimulus. Each cell is one input-output pairing, rows and columns index the source and target models, only models providing embeddings of the required modality are shown per panel. Deeper colors denote higher prediction accuracy, and the direction difference indicates the paired Wilcoxon signed-rank test result of (a) vs. (b). The nonlinear counterpart is reported in Appendix A.9.

chance level. Importantly, our analysis is limited to vision transformers; while convolutional neural networks (CNNs) also capture higher-order semantic information (Yamins et al., 2014), evidence suggests that vision transformers aggregate such global structure earlier and more uniformly across layers (Raghu et al., 2021). Second, residualizing the embeddings shows that a single, removable component, i.e., the information one modality shares with the other, is what drives cross-modal neural encoding performances. In summary, these results point at a directional asymmetry regarding where cross-modal semantic content is encoded in text versus image embeddings.

Nevertheless, our study has several limitations. It relies on a single dataset of eight participants and 252 stimuli, and is based on a subset of most reliable voxels rather than specific regions of interest (ROIs). Moreover, we use linear mappings to quantify the modality-shared part in the residual neural encoding analysis. While our non-linear cross-modal embedding prediction results (see Appendix A.9) do not deviate substantially from the linear case, the role of modality-shared information beyond linear mappings in the context of neural encoding needs to be investigated further. Further research is also needed to specify the exact nature of the modality-shared information across text and image embeddings. Additionally, future work should test the robustness of the asymmetry across *a priori* defined ROIs and extend it to additional modalities and larger participant samples, and ask whether training interventions that narrow the gap between text and image embeddings in VLMs (reviewed in Song et al., 2025) could also improve cross-modal neural encoding performances.

6 Conclusion

Using a paired phrase- and scene-level text-image fMRI dataset, our cross-modal neural encoding approach based on VLM-based text and image embeddings revealed a consistent asymmetry: while text embeddings could predict image-evoked brain activity, image embeddings largely could not predict text-evoked brain activity above chance level. Two additional analyses, namely neural encoding based on residualized embeddings and cross-modal embedding prediction, provided converging evidence that cross-modal information may be more recoverable from text embeddings than from image embeddings, and that image embeddings may contain richer modality-specific information. In the long term, the presented paired text-image neural encoding paradigm could help in training and evaluating image embeddings with the goal of creating representations that capture the same higher-level cross-modal information present in language and the human brain.

Ethics Statement

This study uses a publicly available functional magnetic resonance imaging (fMRI) dataset and text-image dataset; more details on informed consent and institutional ethics approval can be found in the original publication (Balabin et al., 2026). The dataset includes eight participants, who may not represent the broader population, and stimuli with English captions that predominantly depict Western contexts. The evaluated vision-language models and unimodal baseline models likely reflect the social and cultural biases present in their training data, which may in turn influence the measured correspondence between model embeddings and human neural data.

Reproducibility Statement

Both the employed functional magnetic resonance (fMRI) dataset and additional text-image dataset are publicly available. All code used for the analysis is available in an anonymous GitHub repository: `https://anonymous.4open.science/r/cross_modal_neural_encoding_anonymous/README.md`.

AI Use Statement

In this work, besides general phrasing and grammar assistance, we used generative AI tools for the following tasks: provide feedback on research methodology or experiments, implement methods (i.e., generate code with human oversight, specifically for visualizing the results), clean and reformat dataset, create or edit software code, edit a research paper to improve readability. We have not used generative AI tools for the following tasks: help develop theoretical models or conceptual frameworks, propose or refine hypotheses, interpret results, summarize or analyze existing literature, identify relevant literature, format references. The following tasks are not applicable to this work: generate synthetic data sets, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, assist with translation, support qualitative and thematic data analysis. We have reviewed all AI-assisted work, and take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

References

Emily J. Allen, Ghislain St-Yves, Yihan Wu, Jesse L. Breedlove, Jacob S. Prince, Logan T. Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, J. Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, January 2022. ISSN 1546-1726. doi: 10.1038/s41593-021-00962-x. URL `https://www.nature.com/articles/s41593-021-00962-x`.

Richard Antonello, Javier S Turek, Vy Vo, and Alexander Huth. Low-dimensional Structure in the Space of Language Representations is Reflected in Brain Responses. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8332–8344. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper_files/paper/2021/hash/464074179972cbbd75a39abc6954cd12-Abstract.html`.

Richard Antonello, Aditya Vaidya, and Alexander Huth. Scaling laws for language encoding models in fMRI. *Advances in Neural Information Processing Systems*, 36:21895–21907, December 2023. URL `https://papers.nips.cc/paper_files/paper/2023/hash/4533e4a352440a32558c1c227602c323-Abstract-Conference.html`.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629, Vancouver, BC, Canada, June 2023. IEEE. ISBN 979-8-3503-0129-8. doi: 10.1109/CVPR52729.2023.01499. URL `https://ieeexplore.ieee.org/document/10205476/`.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei

Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL Technical Report, November 2025. URL http://arxiv.org/abs/2511.21631. arXiv:2511.21631 [cs].

Helena Balabin, Antonietta Gabriella Liuzzi, Kevin Statz, Marie-Francine Moens, Patrick Dupont, and Rik Vandenberghe. Compositional Complexity in Text and Images. *Neurobiology of Language*, pp. 1–26, May 2026. ISSN 2641-4368. doi: 10.1162/NOL.a.271. URL https://doi.org/10.1162/NOL.a.271.

Lawrence W. Barsalou. Grounded Cognition. *Annual Review of Psychology*, 59(Volume 59, 2008): 617–645, January 2008. ISSN 0066-4308, 1545-2085. doi: 10.1146/annurev.psych.59.103006.093639. URL https://www.annualreviews.org/content/journals/10.1146/annurev.psych.59.103006.093639.

Anna Bavaresco, Ina Klaric, Raquel Fernandez, and Marie-Francine Moens. What Makes Linguistic Representations Good Models of High-Level Visual Perception in the Human Brain?, June 2026a. URL https://annabavaresco.github.io/assets/pdfs/Visual_Semantics_Arxiv.pdf.

Anna Bavaresco, Marianne De Heer Kloots, Sandro Pezzelle, and Raquel Fernández. Vision-Language Models Align with Human Neural Representations in Concept Processing. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3255–3274, Rabat, Morocco, March 2026b. Association for Computational Linguistics. ISBN 979-8-89176-380-7. URL https://aclanthology.org/2026.eacl-long.150/.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling, May 2023. URL http://arxiv.org/abs/2304.01373. arXiv:2304.01373 [cs.CL].

Jeffrey R. Binder, Rutvik H. Desai, William W. Graves, and Lisa L. Conant. Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex (New York, N.Y.: 1991)*, 19(12):2767–2796, December 2009. ISSN 1460-2199. doi: 10.1093/cercor/bhp055.

Michael F. Bonner and Russell A. Epstein. Computational mechanisms underlying cortical responses to the affordance properties of visual scenes. *PLoS Computational Biology*, 14(4):e1006111, April 2018. ISSN 1553-734X. doi: 10.1371/journal.pcbi.1006111. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC5933806/.

Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1):1–10, February 2022. ISSN 2399-3642. doi: 10.1038/s42003-022-03036-1. URL https://www.nature.com/articles/s42003-022-03036-1.

David M. Chan, Rodolfo Corona, Joonyong Park, Cheol Jun Cho, Yutong Bai, and Trevor Darrell. Analyzing The Language of Visual Tokens, November 2024. URL http://arxiv.org/abs/2411.05001. arXiv:2411.05001 [cs.CV].

Nadine Chang, John A. Pyles, Austin Marcus, Abhinav Gupta, Michael J. Tarr, and Elissa M. Aminoff. BOLD5000, a public fMRI dataset while viewing 5000 visual images. *Scientific Data*, 6(1):49, May 2019. ISSN 2052-4463. doi: 10.1038/s41597-019-0052-3. URL https://www.nature.com/articles/s41597-019-0052-3.

Antonio R. Damasio. The Brain Binds Entities and Events by Multiregional Activation from Convergence Zones. *Neural Computation*, 1(1):123–132, March 1989. ISSN 0899-7667. doi: 10.1162/neco.1989.1.1.123. URL https://ieeexplore.ieee.org/abstract/document/6795614.

Adrien Doerig, Tim C. Kietzmann, Emily Allen, Yihan Wu, Thomas Naselaris, Kendrick Kay, and Ian Charest. High-level visual representations in the human brain are aligned with large language models. *Nature Machine Intelligence*, 7(8):1220–1234, August 2025. ISSN 2522-5839. doi: 10.1038/s42256-025-01072-0. URL https://www.nature.com/articles/s42256-025-01072-0.

Oscar Esteban, Christopher J. Markiewicz, Ross W. Blair, Craig A. Moodie, A. Ilkay Isik, Asier Erramuzpe, James D. Kent, Mathias Goncalves, Elizabeth DuPre, Madeleine Snyder, Hiroyuki Oya, Satrajit S. Ghosh, Jessey Wright, Joke Durnez, Russell A. Poldrack, and Krzysztof J. Gorgolewski. fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, 16(1):111–116, January 2019. ISSN 1548-7105. doi: 10.1038/s41592-018-0235-4. URL https://www.nature.com/articles/s41592-018-0235-4.

Umut Güçlü and Marcel A. J. van Gerven. Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *Journal of Neuroscience*, 35(27):10005–10014, July 2015. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.5023-14.2015. URL https://www.jneurosci.org/content/35/27/10005.

Michael Hanke, Florian J. Baumgartner, Pierre Ibe, Falko R. Kaule, Stefan Pollmann, Oliver Speck, Wolf Zinke, and Jörg Stadler. A high-resolution 7-Tesla fMRI dataset from complex natural stimulation with an audio movie. *Scientific Data*, 1:140003, 2014. ISSN 2052-4463. doi: 10.1038/sdata.2014.3.

Alexander G. Huth, Wendy A. de Heer, Thomas L. Griffiths, Frédéric E. Theunissen, and Jack L. Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458, April 2016. ISSN 1476-4687. doi: 10.1038/nature17637. URL https://www.nature.com/articles/nature17637.

Gabriel Ilharco, Mitchell Wortsman, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP, July 2021. URL https://zenodo.org/record/5143773.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision*, 123(1):32–73, May 2017. ISSN 1573-1405. doi: 10.1007/s11263-016-0981-7. URL https://doi.org/10.1007/s11263-016-0981-7.

Matthew A. Lambon Ralph, Elizabeth Jefferies, Karalyn Patterson, and Timothy T. Rogers. The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, 18(1):42–55, January 2017. ISSN 1471-0048. doi: 10.1038/nrn.2016.150. URL https://www.nature.com/articles/nrn.2016.150.

Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. A Survey of State of the Art Large Vision Language Models: Alignment, Benchmark, Evaluations and Challenges. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1578–1597, June 2025. doi: 10.1109/CVPRW67362.2025.00147. URL https://ieeexplore.ieee.org/document/11147946. ISSN: 2160-7516.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1. doi: 10.1007/978-3-319-10602-1_48.

Tom M. Mitchell, Svetlana V. Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L. Malave, Robert A. Mason, and Marcel Adam Just. Predicting human brain activity associated with the

meanings of nouns. *Science (New York, N.Y.)*, 320(5880):1191–1195, May 2008. ISSN 1095-9203. doi: 10.1126/science.1152876.

Thomas Naselaris, Kendrick N. Kay, Shinji Nishimoto, and Jack L. Gallant. Encoding and decoding in fMRI. *NeuroImage*, 56(2):400–410, May 2011. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2010.07.073. URL https://www.sciencedirect.com/science/article/pii/S1053811910010657.

Subba Reddy Oota, Jashn Arora, Vijay Rowtula, Manish Gupta, and Raju S. Bapi. Visio-Linguistic Brain Encoding. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (eds.), *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 116–133, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.11/.

Subba Reddy Oota, Khushbu Pahwa, Mounika Marreddy, Maneesh Kumar Singh, Manish Gupta, and Bapi Raju Surampudi. Multi-modal brain encoding models for multi-modal stimuli. In *Proceedings of the Thirteenth International Conference on Learning Representations*, April 2025. URL https://openreview.net/forum?id=0dELcFHig2.

Subbareddy Oota, Manish Gupta, and Mariya Toneva. Joint processing of linguistic properties in brains and language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 18001–18014, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/3a0e2de215bd17c39ad08ba1d16c1b12-Abstract-Conference.html.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision, February 2024. URL http://arxiv.org/abs/2304.07193. arXiv:2304.07193 [cs].

Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J. Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, 9(1):963, March 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-03068-4. URL https://www.nature.com/articles/s41467-018-03068-4.

Sara F. Popham, Alexander G. Huth, Natalia Y. Bilenko, Fatma Deniz, James S. Gao, Anwar O. Nunez-Elizalde, and Jack L. Gallant. Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nature Neuroscience*, 24(11):1628–1636, November 2021. ISSN 1546-1726. doi: 10.1038/s41593-021-00921-6. URL https://www.nature.com/articles/s41593-021-00921-6.

Jacob S. Prince, John A. Pyles, Michael J. Tarr, and Kendrick N. Kay. GLMsingle: a turnkey solution for accurate single-trial fMRI response estimates. *Journal of Vision*, 21(9):2831, September 2021. ISSN 1534-7362. doi: 10.1167/jov.21.9.2831. URL https://doi.org/10.1167/jov.21.9.2831.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do Vision Transformers See Like Convolutional Neural Networks? In *Advances in Neural Information Processing Systems*, volume 34, pp. 12116–12128. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/652cf38361a209088302ba2b8b7f51e0-Abstract.html.

Ariel Rokem and Kendrick Kay. Fractional ridge regression: a fast, interpretable reparameterization of ridge regression. *GigaScience*, 9(12):giaa133, November 2020. ISSN 2047-217X. doi: 10.1093/gigascience/giaa133. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC7702219/.

Matteo Ruggero Ronchi and Pietro Perona. Describing Common Human Visual Actions in Images, June 2015. URL http://arxiv.org/abs/1506.02203. arXiv:1506.02203 [cs].

Shreya Saha, Shurui Li, Greta Tuckute, Yuanning Li, Ru-Yuan Zhang, Leila Wehbe, Evelina Fedorenko, and Meenakshi Khosla. Modeling the language cortex with form-independent and enriched representations of sentence meaning reveals remarkable semantic abstractness, September 2025. URL http://arxiv.org/abs/2510.02354. arXiv:2510.02354 [cs].

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45), November 2021. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2105646118. URL https://www.pnas.org/content/118/45/e2105646118.

Simon Schrodi, David T. Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. Two Effects, One Trigger: On the Modality Gap, Object Bias, and Information Imbalance in Contrastive Vision-Language Models. In *The Thirteenth International Conference on Learning Representations*, October 2024. URL https://openreview.net/forum?id=uAFHCZRmXk.

Riki Shimizu, Richard J. Antonello, Chandan Singh, and Nima Mesgarani. Interpretable Embeddings of Speech Enhance and Explain Brain Encoding Performance of Audio Models, September 2025. URL http://arxiv.org/abs/2507.16080. arXiv:2507.16080 [q-bio.NC].

Shezheng Song, Xiaopeng Li, Shasha Li, Shan Zhao, Jie Yu, Jun Ma, Xiaoguang Mao, Weimin Zhang, and Meng Wang. How to Bridge the Gap Between Modalities: Survey on Multimodal Large Language Model. *IEEE Transactions on Knowledge and Data Engineering*, 37(9):5311–5329, September 2025. ISSN 1558-2191. doi: 10.1109/TKDE.2025.3527978. URL https://ieeexplore.ieee.org/abstract/document/10841938.

Johannes Stelzer, Yi Chen, and Robert Turner. Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): random permutations and cluster size control. *NeuroImage*, 65:69–82, January 2013. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2012.09.063.

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 *Advances in Neural Information Processing Systems*, volume 36, pp. 29654–29666, December 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/5ebbbac62b968254093023f1c95015d3-Abstract-Conference.html.

Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Proceedings of the 33rd Conference on Neural Information Processing Systems*, Vancouver, Canada, November 2019. URL http://arxiv.org/abs/1905.11833. arXiv: 1905.11833.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency, August 2025. URL http://arxiv.org/abs/2508.18265. arXiv:2508.18265 [cs].

Haiguang Wen, Junxing Shi, Yizhen Zhang, Kun-Han Lu, Jiayue Cao, and Zhongming Liu. Neural Encoding and Decoding with Deep Learning for Dynamic Natural Vision. *Cerebral Cortex*, 28(12):4136–4160, December 2018. ISSN 1047-3211. doi: 10.1093/cercor/bhx268. URL https://doi.org/10.1093/cercor/bhx268.

Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, June 2014. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1403112111. URL https://www.pnas.org/content/111/23/8619.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: Open Pre-trained Transformer Language Models, June 2022. URL http://arxiv.org/abs/2205.01068. arXiv:2205.01068 [cs].

A Appendix

A.1 Layer Tests

To examine the effect of the chosen layer within the investigated vision-language models that we based our main analysis on, we repeated the main neural encoding analysis at every layer within Qwen3.5-VL 9B. Vision-encoder performance rises steeply over the first ten blocks and then plateaus (image→image: 0.033 at block 0, 0.086 at block 13, peak 0.088 at block 20), whereas the text encoder plateaus across states 3–16 before declining over the final third of the stack (text→text: 0.032 at state 14, falling to 0.018 at state 32). In every condition the middle layer closely approaches the performance of the best performing layer, making it a reliable estimate of overall neural encoding performances.

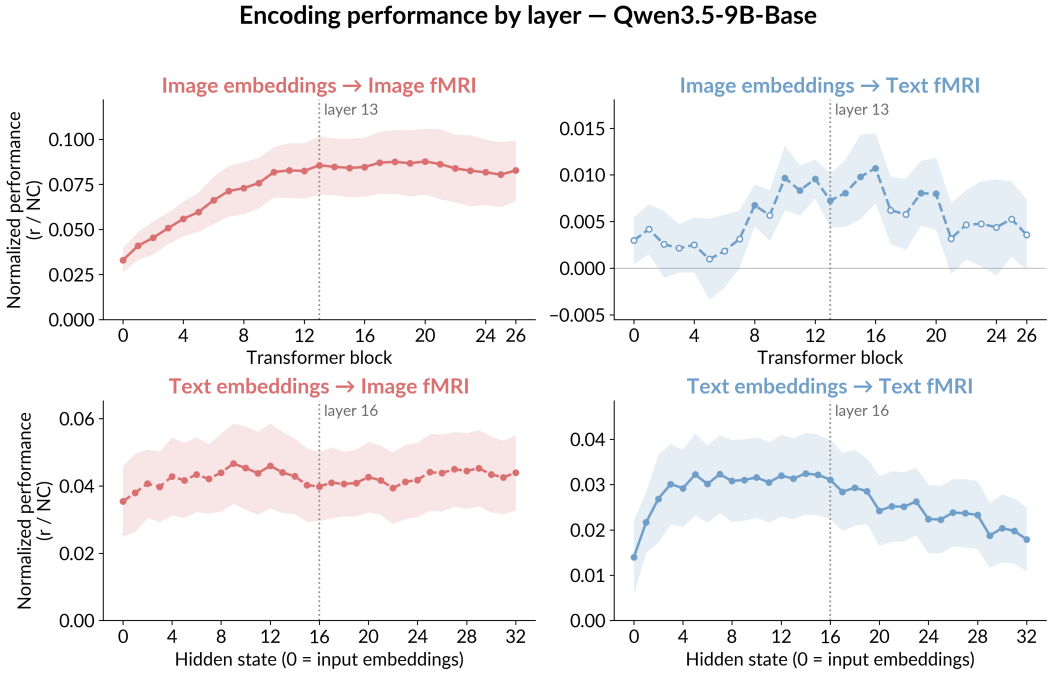


Figure 5: **Encoding performance as a function of layer depth within Qwen3.5-VL 9B.** Each panel shows one condition, the depicted curves show the group mean of the noise-ceiling-normalized encoding correlation across the eight participants, with shaded ± standard error. Filled markers denote layers significant against the permutation null at $q < 0.05$ (corrected across all layer × condition tests), open markers non-significant; the dotted rule marks the middle layer used in all other analyses.

A.2 Noise Ceiling Estimation

Trial-wise neural responses (i.e., betas) are z-scored within each run, so that the total variance of each voxel is equal to one. For each voxel, the noise variance σ^2_{noise} is the trial-to-trial variance of the betas across repeats of the same stimulus, averaged over stimuli, the signal variance is $\sigma^2_{\text{signal}} = \max(1 - \sigma^2_{\text{noise}}, 0)$. The noise-ceiling (NC) signal-to-noise ratio (SNR) is then $\text{NCSNR} = \sigma_{\text{signal}}/\sigma_{\text{noise}}$, converted to a noise-ceiling correlation $r_{\text{NC}} = \sqrt{\text{NC}\%/100}$, where $\text{NC}\% = 100 \cdot \text{NCSNR}^2/(\text{NCSNR}^2 + 1/n_{\text{avg}})$ with $n_{\text{avg}} = 6$ (in line with six repetitions per stimulus, see §3.1). Noise ceilings are computed separately for image- and text-evoked responses to reflect the modality-specific reliability of each condition. The primary reported metric is the noise-ceiling-normalized accuracy $r_{\text{norm}} = r/r_{\text{NC}}$, computed over the top 20% of voxels by noise ceiling within each modality. The spatial distribution of the top 20% selected voxels per subject are shown in Appendix A.5. Additional replications of the main neural encoding analysis (RQ1) with a larger set of voxels (i.e., 40% and 60% of the voxels with the highest noise ceiling) are provided in Appendix A.3.

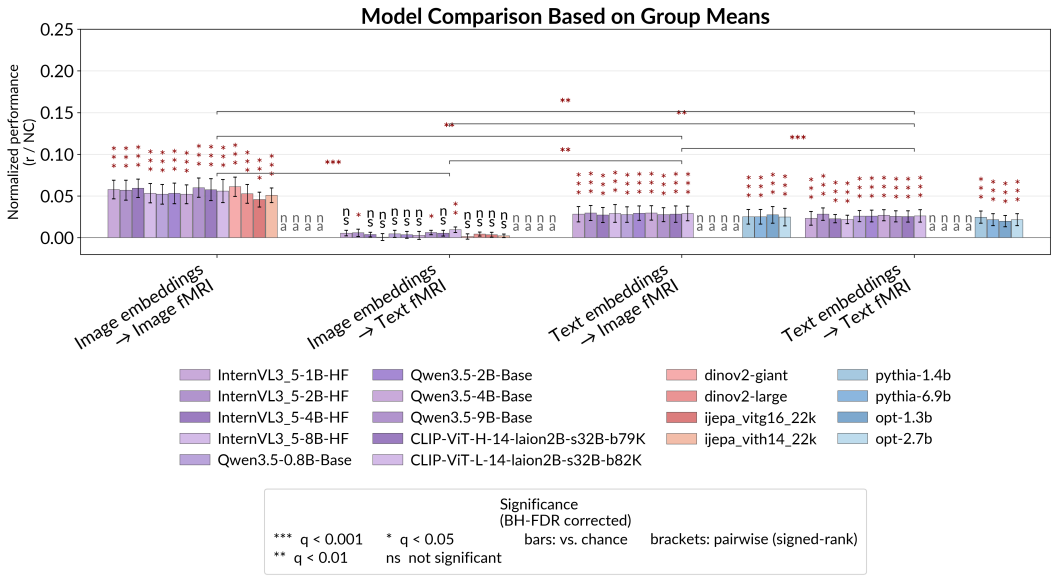


Figure 6: **Replicated cross-modal neural encoding analysis (RQ1)**, but with the top 40% most reliable voxels per modality instead of the top 20%. For more details on the figure annotations, see Figure 2.

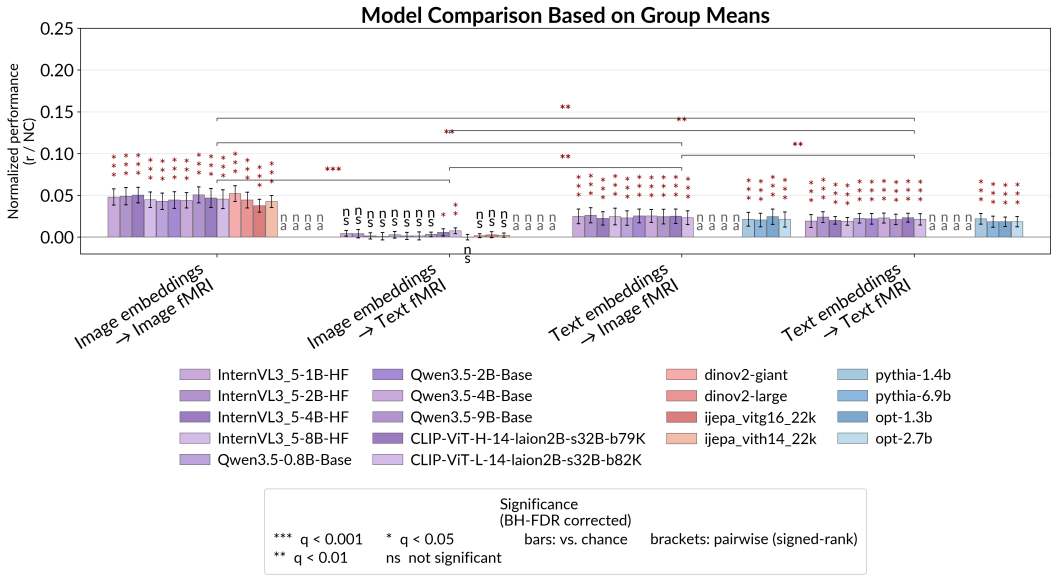


Figure 7: **Replicated cross-modal neural encoding analysis (RQ1)**, but with the top 60% most reliable voxels per modality instead of the top 20%. For more details on the figure annotations, see Figure 2.

A.3 Top % of Voxels

As described in Appendix A.2, the main analysis retained the 20% most reliable voxels per fMRI modality. To investigate how the choice of threshold for the most reliable voxels affects the results, we repeated the main neural encoding analysis (RQ1) for the 40% and 60% most reliable voxels, as shown in Figures 6 and 7. Across both additional thresholds, the relative performance differences between the four tested conditions were preserved, indicating that the reported effects do not depend

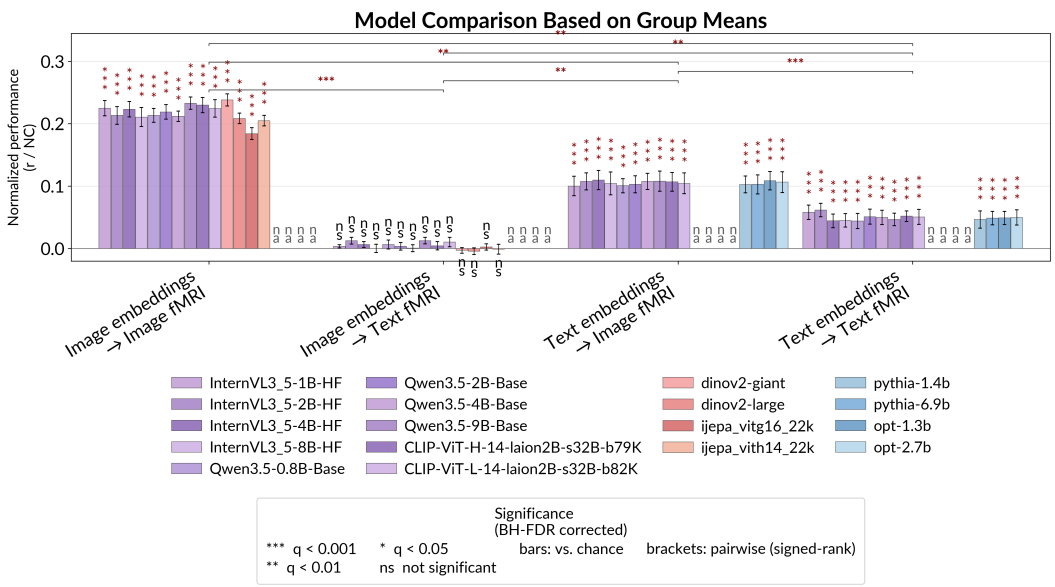


Figure 8: **Cross-modal neural encoding analysis with a shared set of voxels across image-based and text-based fMRI responses (RQ1)**. For more details on the figure annotations, see Figure 2.

on the choice of threshold. Nonetheless, absolute performances were lower with an increasing number of included voxels, indicating that a higher number of voxels may introduce additional noise.

A.4 Intersection of Text- and Vision-Specific Voxels

In the main analysis, the top 20% of voxels with the highest noise ceiling were selected per fMRI modality, so image and text conditions were evaluated on partly non-overlapping voxels. The observed cross-modal asymmetry might thus be potentially biased by the selection of image- or text-specific voxels. To rule out this confound, we repeated the analysis on a single voxel set shared by all conditions in the 2×2 experimental setup, defined per subject as the intersection of the top 20% image-based noise ceiling voxels and top 20% text-based noise-ceiling voxels. All other procedures were unchanged, including noise-ceiling normalization by each fMRI modality's own ceiling. Figure 8 replicates the main finding: text embeddings reliably predict image-evoked brain activity, whereas image embeddings are unable to predict text-evoked brain activity above chance level.

A.5 Brain Maps of Voxels with the Highest Noise Ceiling

Figures 9–16 show, for each participant, the spatial distribution of the 20% most reliable voxels, alongside the top 40% and 60% used in the additional analyses included in Appendix A.3.

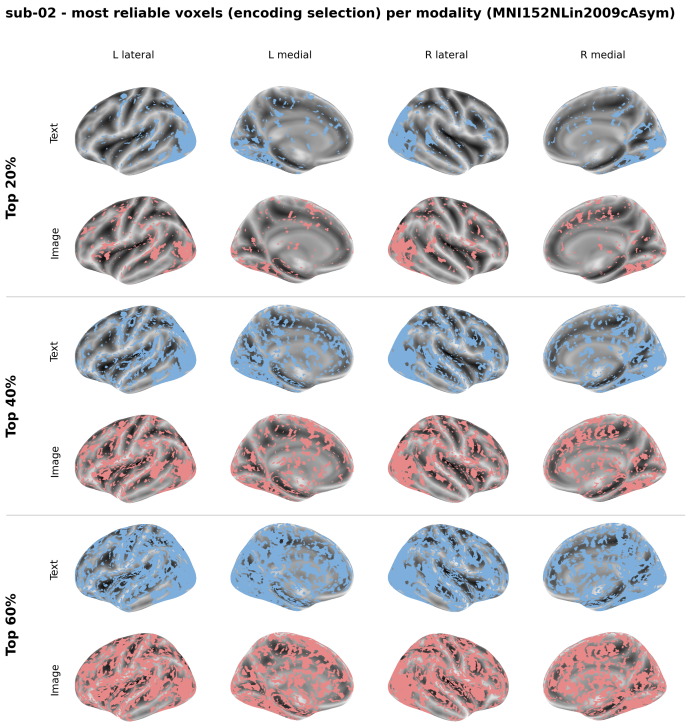


Figure 9: Most reliable voxels in sub-02, per modality, at three selection levels (top 20/40/60% of positive-noise-ceiling voxels; text blue, image red). Noise ceiling estimation and voxel selection are done in native functional volume space. For visualization purposes, the voxels are warped to MNI152NLin2009cAsym and sampled onto fsaverage6. L; left, R; right.

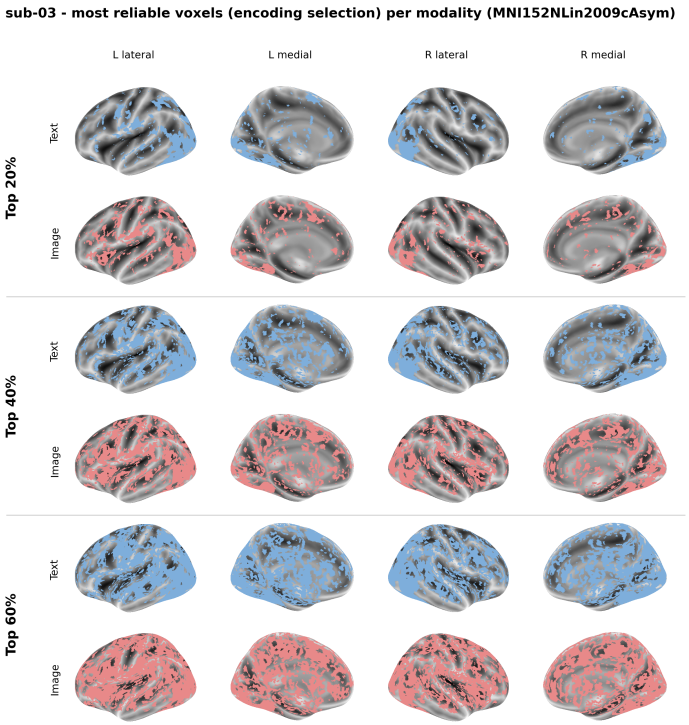


Figure 10: Most reliable voxels in sub-03, per modality, at three selection levels. Conventions as in Figure 9.

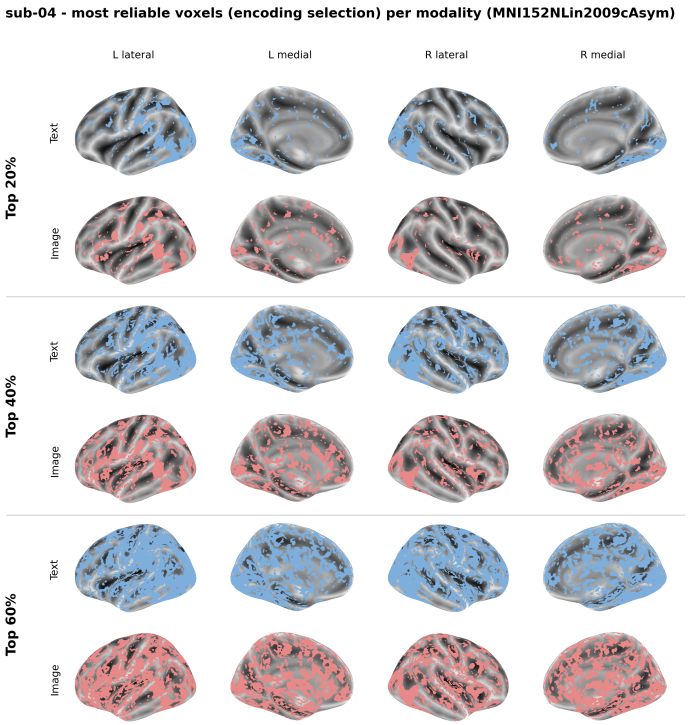


Figure 11: Most reliable voxels in sub-04, per modality, at three selection levels. Conventions as in Figure 9.

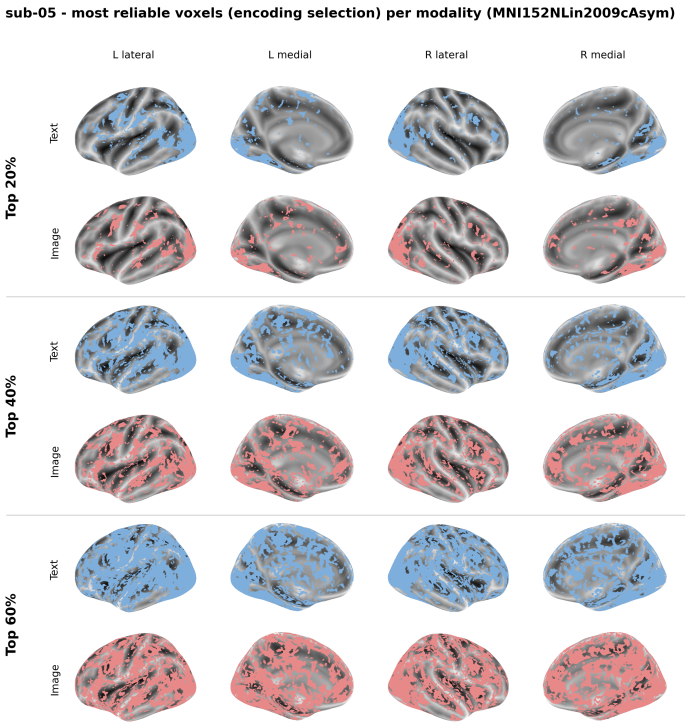


Figure 12: Most reliable voxels in sub-05, per modality, at three selection levels. Conventions as in Figure 9.

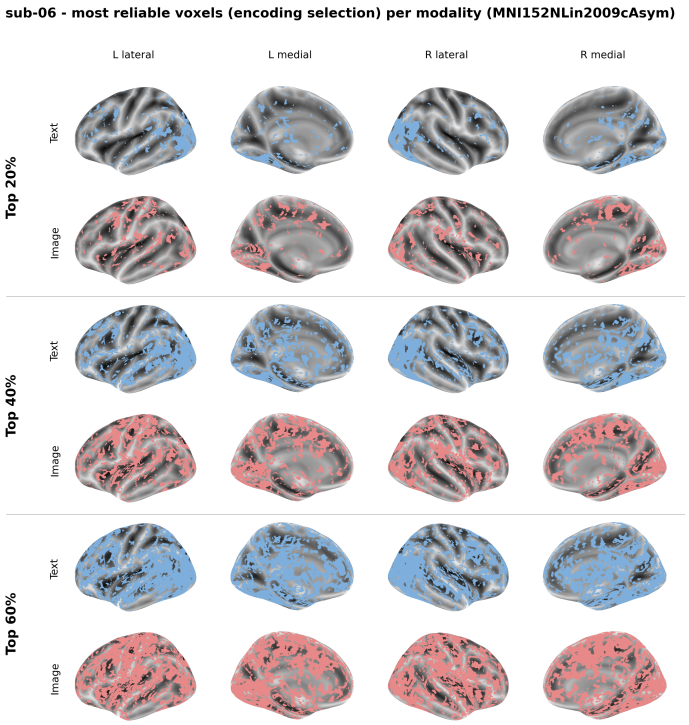


Figure 13: Most reliable voxels in sub-06, per modality, at three selection levels. Conventions as in Figure 9.

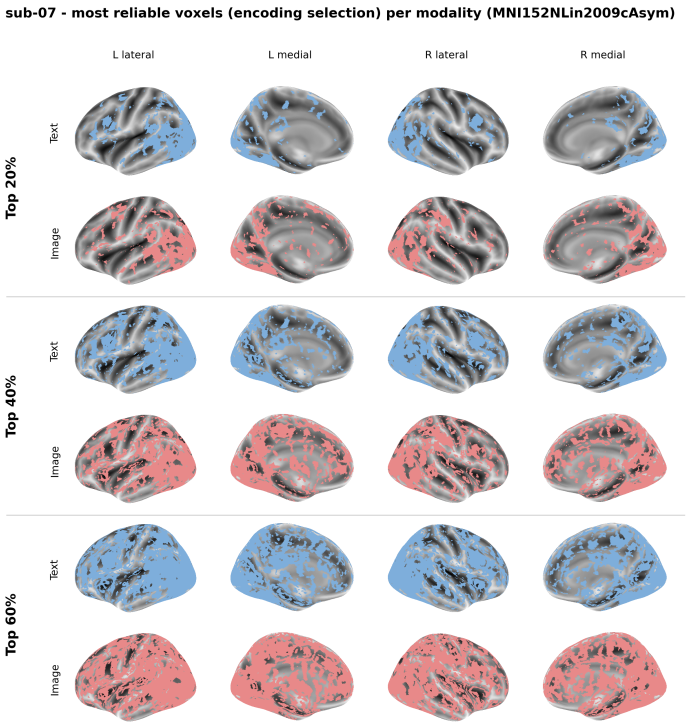


Figure 14: Most reliable voxels in sub-07, per modality, at three selection levels. Conventions as in Figure 9.

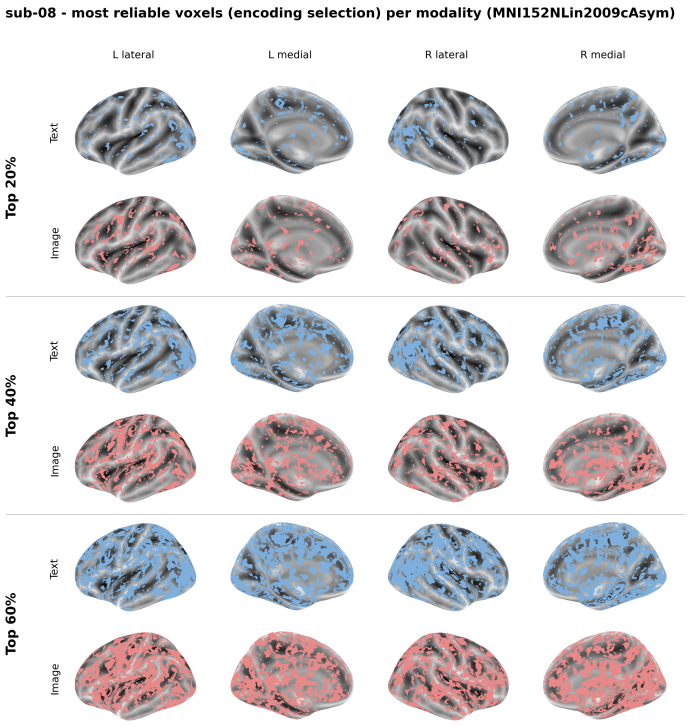


Figure 15: Most reliable voxels in sub-08, per modality, at three selection levels. Conventions as in Figure 9.

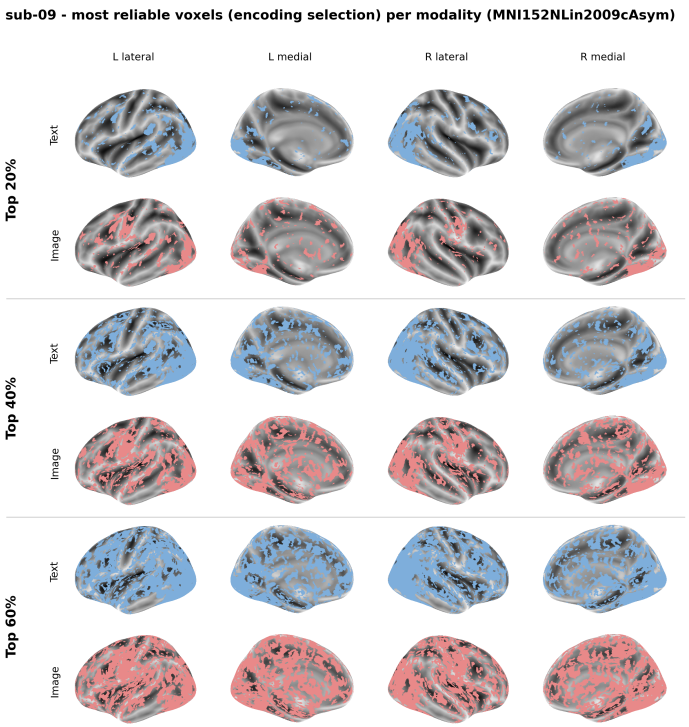


Figure 16: Most reliable voxels in sub-09, per modality, at three selection levels. Conventions as in Figure 9.

A.6 Subject-Wise Neural Encoding Performances

Figure 17 shows the encoding accuracies reported in Section §4.1 by subject rather than by model, averaged across all evaluated models per subject within each of the four conditions. The asymmetry described in the main text holds at the individual level as well: text embeddings predict image-evoked fMRI activity in every subject, whereas image embeddings do not predict text-evoked fMRI activity above chance level in any subject, indicating that the group-level effect is not driven by a subset of participants.

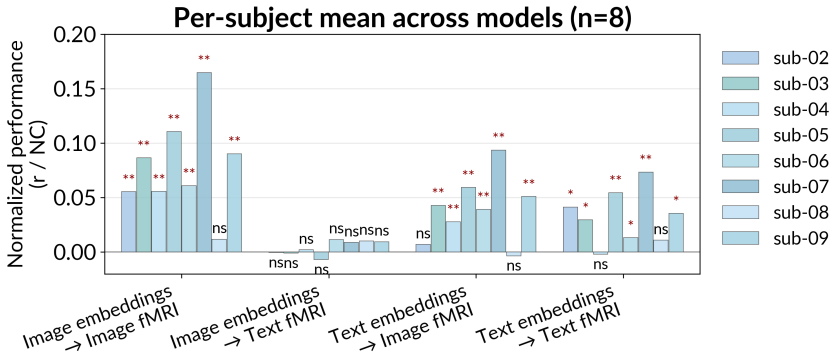


Figure 17: **Subject-wise cross-modal neural encoding (RQ1).** Noise-ceiling-normalized encoding accuracy (r/r_{NC}, averaged across all tested models, computed over the top 20% of voxels by noise ceiling) shown per subject and per condition.

A.7 fMRI-Based Residual Neural Encoding

The residualization described in §3.4 removes the modality-shared component from the VLM embedding. Here, we apply the analogous manipulation to the brain response itself, thereby examining the role of modality-shared information from the brain rather than the model side.

Setup Let $\mathbf{y}$ be the fMRI response, $\mathbf{e}$ the PCA-reduced embedding used as the encoding input, and $\mathbf{r}$ the PCA-reduced embedding of the modality that is not $\mathbf{e}$. We fit a ridge regression model $\mathbf{W}$ to predict $\mathbf{y}$ from $\mathbf{r}$ on the training stimuli (using a regularization strength of $\alpha = 1.0$), form the prediction $\hat{\mathbf{y}} = \mathbf{W}\mathbf{r}$, and keep only the residual $\tilde{\mathbf{y}}$ it does not explain:

$$\tilde{\mathbf{y}} = \mathbf{y} - \hat{\mathbf{y}}, \qquad \hat{\mathbf{y}} = \mathbf{W}\mathbf{r}, \qquad \mathbf{W} = \arg\min_{\mathbf{W}} \sum_{i \in \text{train}} \left\| \mathbf{y}_i - \mathbf{W}\mathbf{r}_i \right\|_2^2 + \alpha \|\mathbf{W}\|_F^2 \tag{2}$$

The unmodified embedding $\mathbf{e}$ is then used to predict $\tilde{\mathbf{y}}$. Because $\mathbf{r}$ is formed from the embeddings, the manipulation is asymmetric with respect to the embedding-based variant; all four conditions of the 2×2 design are therefore listed explicitly in Table 2. Statistical testing follows §3.4.

Condition		Prediction	Residual fMRI target $\tilde{\mathbf{y}}$	Encode → Target
text→text	(within)	$\hat{\mathbf{y}}_{\text{text}} = \mathbf{W}\mathbf{e}_{\text{image}}$	$\tilde{\mathbf{y}}_{\text{text}} = \mathbf{y}_{\text{text}} - \hat{\mathbf{y}}_{\text{text}}$	$\mathbf{e}_{\text{text}} \rightarrow \tilde{\mathbf{y}}_{\text{text}}$
image→image	(within)	$\hat{\mathbf{y}}_{\text{image}} = \mathbf{W}\mathbf{e}_{\text{text}}$	$\tilde{\mathbf{y}}_{\text{image}} = \mathbf{y}_{\text{image}} - \hat{\mathbf{y}}_{\text{image}}$	$\mathbf{e}_{\text{image}} \rightarrow \tilde{\mathbf{y}}_{\text{image}}$
image → text	(cross)	$\hat{\mathbf{y}}_{\text{text}} = \mathbf{W}\mathbf{e}_{\text{text}}$	$\tilde{\mathbf{y}}_{\text{text}} = \mathbf{y}_{\text{text}} - \hat{\mathbf{y}}_{\text{text}}$	$\mathbf{e}_{\text{image}} \rightarrow \tilde{\mathbf{y}}_{\text{text}}$
text → image	(cross)	$\hat{\mathbf{y}}_{\text{image}} = \mathbf{W}\mathbf{e}_{\text{image}}$	$\tilde{\mathbf{y}}_{\text{image}} = \mathbf{y}_{\text{image}} - \hat{\mathbf{y}}_{\text{image}}$	$\mathbf{e}_{\text{text}} \rightarrow \tilde{\mathbf{y}}_{\text{image}}$

Table 2: **fMRI-based residualization conditions.** As in Table 1, but the modality-shared component is removed from the brain response rather than from the embedding. The prediction $\hat{\mathbf{y}}$ is obtained by ridge-regressing the fMRI response $\mathbf{y}$ on the PCA-reduced embedding $\mathbf{r}$ of the modality that is not the encoding input, fit on the training stimuli only; the residual target $\tilde{\mathbf{y}} = \mathbf{y} - \hat{\mathbf{y}}$ is then predicted from the unmodified embedding $\mathbf{e}$. Since $\mathbf{r}$ is defined relative to the encoding input, conditions sharing an fMRI target are residualized against different regressors, rendering this manipulation asymmetric with respect to its embedding-based counterpart.

Results Residualizing the neural responses against the embedding of the other modality reproduces the pattern observed on the embedding side. As shown in Figures 18- 20, residualized representations yield significantly larger decreases in neural encoding performance in the cross-modal compared to the within-modality setup. As on the embedding side, we report the two complementary views of the same analysis: Figure 19 shows the corresponding absolute differences ($\Delta\, r/r_{\mathrm{NC}}$ = residualized − baseline), and Figure 20 the residualized accuracies (r/r_{NC}) from which they are computed. Taken together with §4.2, removing the shared component on either the model or the brain side abolishes the same cross-modal effect.

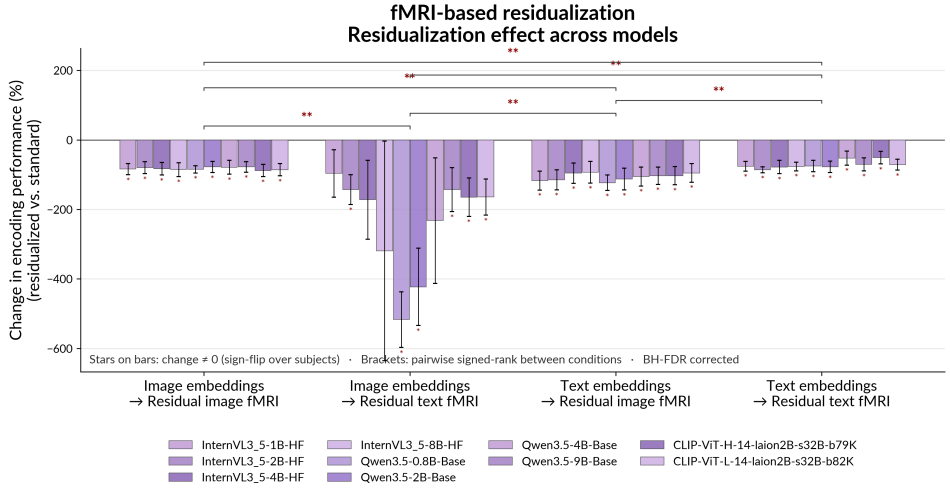


Figure 18: **fMRI-based residual neural encoding (RQ2).** Relative change in noise-ceiling-normalized encoding accuracy (%Δ, computed with respect to the unablated baseline) when the component linearly predictable from the embedding of the respective other modality is removed from the fMRI response, following the residualization setups listed in Table 2, shown per model (colored by family) and per condition; error bars denote across-subject variability. Negative values indicate a loss of encoding accuracy relative to the unablated baseline. Significance annotations follow Figure 3.

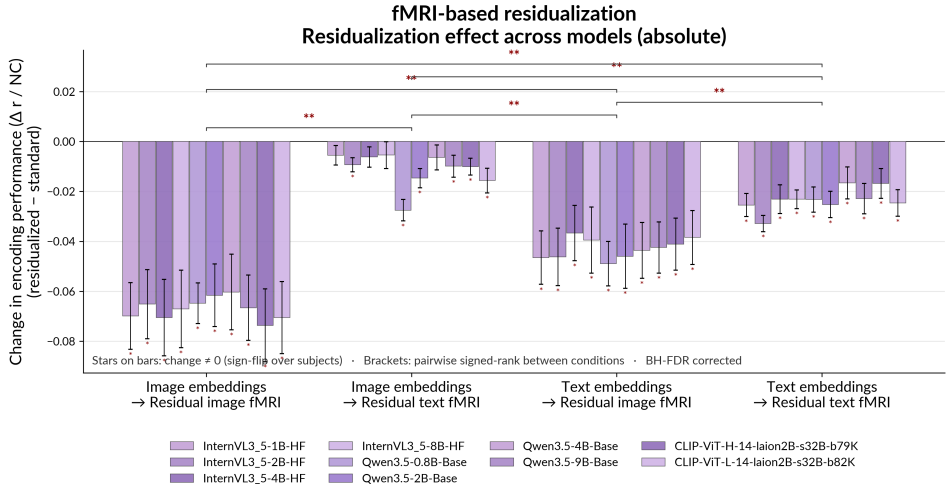


Figure 19: **fMRI-based residual neural encoding (RQ2).** Absolute difference in noise-ceiling-normalized encoding accuracy ($\Delta\, r/r_{\mathrm{NC}}$ = residualized − baseline) when the component linearly predictable from the embedding of the respective other modality is removed from the fMRI response, following the residualization setups listed in Table 2, shown per model (colored by family) and per condition; error bars denote across-subject variability. Significance annotations follow Figure 21.

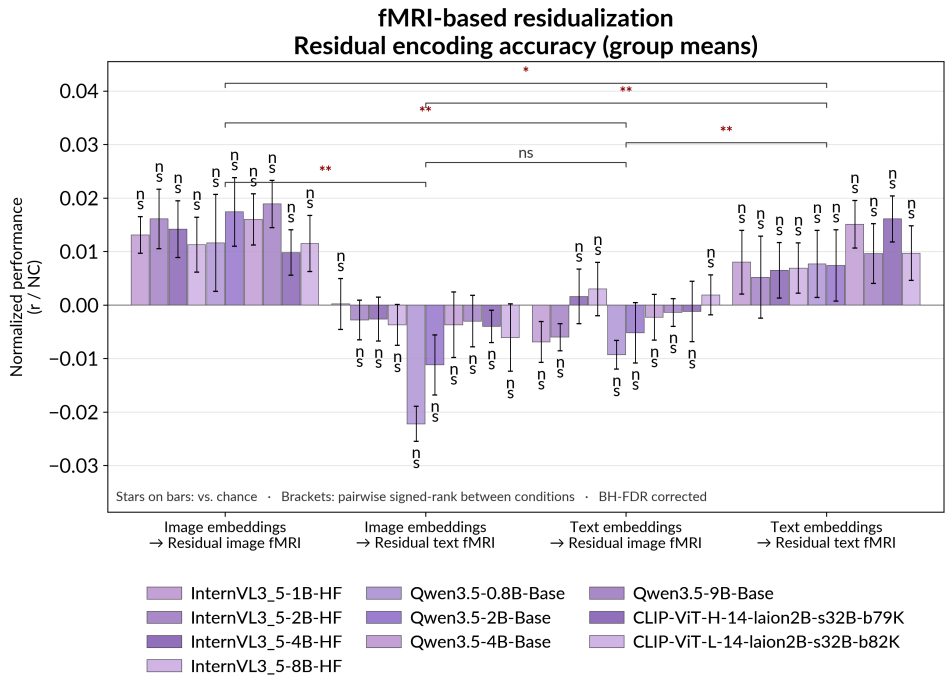


Figure 20: **fMRI-based residual neural encoding (RQ2).** Residualized noise-ceiling-normalized encoding accuracy (r/r_{NC}, group mean over the eight subjects, computed over the top 20% of voxels by noise ceiling) obtained after the component linearly predictable from the embedding of the respective other modality is removed from the fMRI response, following the residualization setups listed in Table 2, shown per model (colored by family) and per condition; error bars denote across-subject variability. These are the absolute accuracies from which the differences in Figure 19 are computed. Significance levels are assessed in the same manner as in Figure 2.

A.8 Absolute and Relative Differences Between Baseline and Residualized Neural Encoding Models

Figure 3 in the main text of the manuscript reports the effect of the embedding-based residualization based on the absolute neural encoding scores. For completeness, we report two complementary views of the same analysis here. Figure 21 shows the absolute difference ($\Delta\, r/r_{\mathrm{NC}}$ = residualized − baseline), expressed in the same units as the encoding accuracies themselves, and Figure 22 shows the relative difference between the baseline and residualized accuracies.

A.9 Non-Linear Cross-Modal Embedding Prediction

The Pearson correlations of all pairwise non-linear cross-modal embedding predictions are shown in Figure 23, based on a four-layer skip-connected multilayer perceptron (hidden width 256, rectified linear unit (ReLU) activations, trained with Adam at a learning rate of 10^{-3} for up to 256 epochs with early stopping).

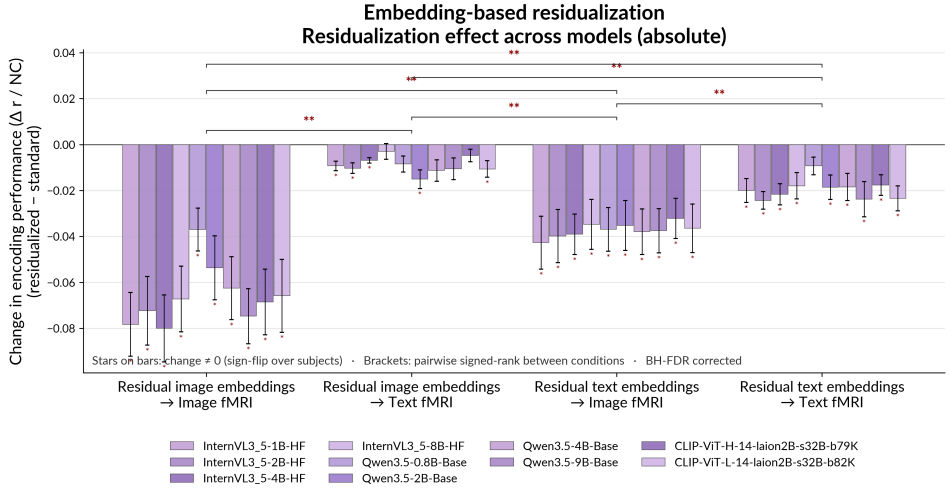


Figure 21: **Embedding-based residual neural encoding (RQ2).** Absolute difference in noise-ceiling-normalized encoding accuracy ($\Delta\, r / r_{\mathrm{NC}}$ = residualized − baseline) when the component of each embedding linearly predictable from the other modality is removed, following the residualization setups listed in Table 1, shown per model (colored by family) and per condition; error bars denote across-subject variability. Stars on bars denote a change significantly different from zero (sign-flip test across subjects), and brackets denote pairwise signed-rank tests between conditions, both BH-FDR corrected.

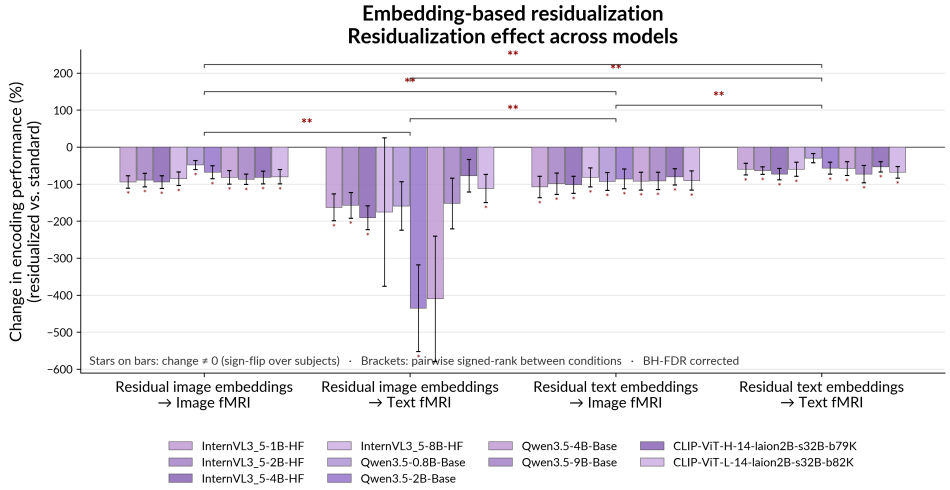
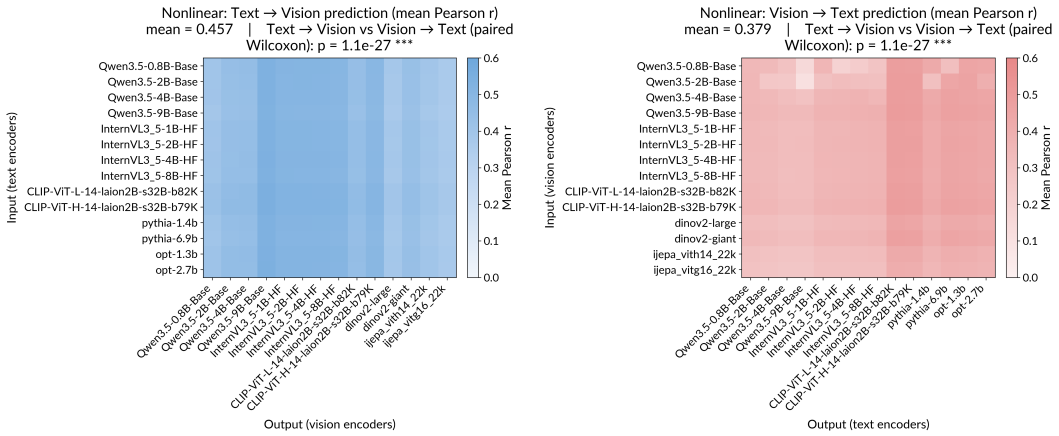


Figure 22: **Embedding-based residual neural encoding (RQ2).** Relative difference in noise-ceiling-normalized encoding accuracy ($\%\Delta$ computed with respect to the unablated baseline) when the component of each embedding linearly predictable from the other modality is removed, following the residualization setups listed in Table 1, shown per model (colored by family) and per condition; error bars denote across-subject variability. Stars on bars denote a change significantly different from zero (sign-flip test across subjects), and brackets denote pairwise signed-rank tests between conditions, both BH-FDR corrected.



(a) Text-to-image: predicting the pooled vision embedding of the target model (columns) from the pooled text embedding of the source model (rows).

(b) Image-to-text: predicting the pooled text embedding of the target model (columns) from the pooled vision embedding of the source model (rows).

Figure 23: **Non-linear cross-modal embedding prediction (RQ3).** Pearson r of non-linearly predicting the pooled embedding of one modality from the pooled embedding of the other modality for the same stimulus, evaluated on the $46,944$ image-caption pairs of the VG-COCO overlap set. Each cell is one input-output pairing, with rows and columns indexing the source and target models, respectively; only models providing embeddings of the required modality are shown per panel. Bolder colors denote higher prediction accuracy.