

See Beyond Words: Joint Visual–Language Representations Improve Brain Encoding of Language-Only Stimuli

Yongpeng Zhang, Congchi Yin, Runze Xia, Xuyun Wen, Piji Li*

College of Artificial Intelligence, Nanjing University of Aeronautics and Astronautics,
Nanjing, China

The Key Laboratory of Brain-Machine Intelligence Technology, Ministry of Education,
Nanjing, China

Email: {zhangyongpeng, pjli}@nuaa.edu.cn

Abstract

Human language processing in the brain involves the coordinated engagement of multiple perceptual pathways. In parallel, recent multimodal large language models (MLLMs) integrate visual and linguistic information into language generation. However, it remains unclear whether and how the injection of visual information alters language representations, and how such changes in turn affect neural predictions during language processing. To address this question, we analyze joint vision–language representations extracted from MLLMs using brain encoding models. We visually augment textual stimuli via image generation and use the resulting representations to predict fMRI responses recorded while participants process language-only stimuli. We find that joint vision–language representations substantially improve neural prediction performance, with pronounced gains in higher-level visual regions. Further analyses indicate that these improvements reflect both the engagement of visual pathways and the enrichment of linguistic representations through visually grounded information. Together, our results provide evidence for multimodal pathway involvement in human language processing and offer insights into the interpretability of multimodal models.

1 Introduction

Brain encoding predicts neural activity from external stimuli representations, offering a powerful framework to study human information processing. Recently, researchers increasingly derive these representations from pretrained deep neural networks across language (Wehbe et al., 2014; Jain and Huth, 2018; Toneva and Wehbe, 2019; Caucheteux and King, 2022), vision (Yamins et al., 2014; Eickenberg et al., 2017; Schrimpf et al., 2018; Wang et al., 2023), and speech (Millet et al., 2022; Vaidya et al.,

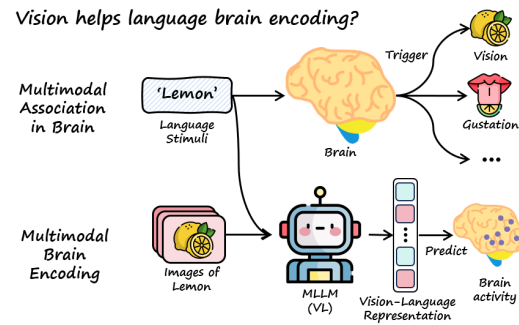


Figure 1: Our motivation: If language processing in the brain evokes visual associations, whether and how the injection of visual information alters language representations, and how such changes in turn affect neural predictions during language processing?

2022; Tuckute et al., 2023), bridging deep learning representations with human semantic spaces.

Human cognition relies on multi-sensory integration; for example, hearing “lemon” evokes both visual and gustatory imagery, aligning with mental imagery theory (Spence and Deroy, 2012; Nanay, 2018). This principle mirrors the evolution of Multimodal Large Language Models (MLLMs), which ground core language capabilities in multi-sensory inputs like vision (Liu et al., 2023; Dubey et al., 2024; Wang et al., 2024). In brain encoding, prior work demonstrates that vision-language model (VLM) representations improve neural activity prediction when participants process *multimodal inputs* (Oota et al., 2025) or *visual stimuli* (Tang et al., 2023; Wang et al., 2023).

This raises an unexplored question: *Can VLM representations of vision-language stimuli better predict neural responses to **language-only stimuli**?* Intuitively, language processing engages multimodal neural pathways. Incorporating visual information into language representations may therefore capture these cross-modal activations more effectively. Investigating this unimodal setting pro-

*Corresponding author

vides crucial evidence for multimodal pathway engagement during language comprehension, yielding new insights into MLLM interpretability. We investigate this by visually augmenting textual stimuli and extracting joint vision-language representations from MLLMs to predict fMRI responses to language. These joint representations substantially outperform language-only baselines in neural prediction, particularly along visual pathways. We analyze these improvements by examining both human brain response organization and MLLM internal representational structures.

Exploration of this setting previously faced two challenges. First, paired multimodal data is asymmetric; while image captioning is common, generating aligned images for continuous text is difficult, making relevant multimodal fMRI datasets scarce. Second, traditional VLMs (e.g., CLIP (Radford et al., 2021), VisualBERT (Li et al., 2019)) optimize for cross-modal matching and perform poorly on unimodal inputs, hindering information flow analysis. We overcome these barriers by using modern image generation to visually augment language stimuli, constructing paired multimodal data. Furthermore, we employ recent MLLMs capable of robustly processing both unimodal and multimodal inputs, enabling controlled brain encoding experiments and deeper mechanistic insights.

Our contributions are threefold:

1. We demonstrate that vision-language representations better predict neural responses to language-only stimuli than unimodal baselines, with pronounced gains in visual pathways.
2. We quantify visual information’s contribution to neural prediction, providing evidence for visual pathway involvement during human language comprehension.
3. We reveal that under explicit visual input, MLLM textual representations distribute more broadly and structurally align with neural representations along visual pathways, illuminating multimodal model mechanisms.

2 Related Work

Early studies of interactions among modality pathways in human cognition primarily adopted cross-modal paradigms, largely due to the lack of effective methods for integrating multimodal information during data acquisition and analysis (Thirion

et al., 2006; Devereux et al., 2013; Popham et al., 2021). With the advent of high-performing multimodal models, an increasing number of studies have leveraged their representations to investigate brain activity. Prior work has shown that multimodal models outperform unimodal ones in predicting neural responses to visual stimuli (Doerig et al., 2022), and that representations of multimodal inputs such as videos can be used to predict brain activity during video viewing (Dong and Toneva, 2023; Oota et al., 2025). Comparisons between multimodal and unimodal models further quantified the contributions of modality-specific information to brain encoding. Closely related to our work, several studies examined whether language information enhances the prediction of neural responses to visual stimuli. For instance, (Wang et al., 2023) used CLIP-based joint image–text representations and demonstrated improved prediction in visual cortical areas. Extending this line of work, (Oota et al., 2022) compared representative multimodal architectures at different fusion stages, including CLIP, VisualBERT, and LXMERT. Notably, these traditional architectures are typically trained with cross-modal alignment objectives, whereas state-of-the-art multimodal models are predominantly trained with generation-based objectives. While the former emphasize explicit cross-modal alignment, the latter integrate multimodal information into language-centric representations, a distinction that closely matches the motivation of our study.

3 Predicting Neural Responses with Vision–Language Joint Representations

We use joint vision–language representations to predict brain activity elicited by language-only stimuli. In this section, we describe our methods and experiments on large-scale datasets using high-capacity MLLMs. Our central question is whether incorporating visual information into language representations improves the prediction of neural responses to language. Unlike prior work that primarily relies on traditional VLMs trained with cross-modal matching objectives, we employ MLLMs that perform robustly in both language-only and multimodal settings. Crucially, visual and linguistic information is processed through a shared backbone, enabling controlled and reliable comparisons across input conditions.

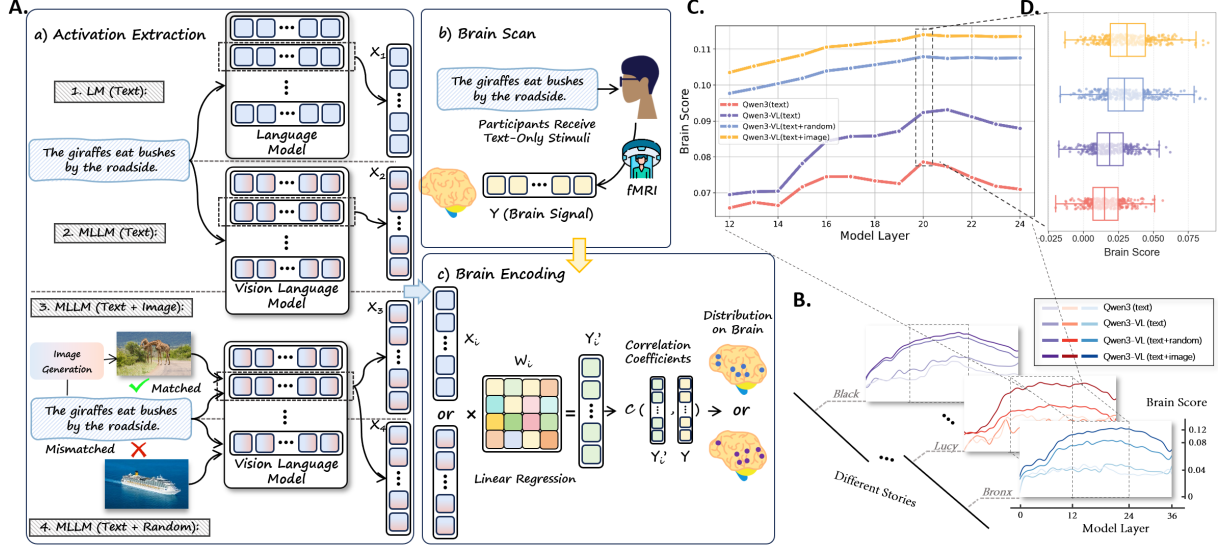


Figure 2: **A** Overview of the experimental pipeline. **A.a** shows four control settings combining unimodal and multimodal models and inputs. Results are illustrated using the Qwen3 family; results for other models are reported in the appendix H. **A.b** depicts the fMRI acquisition procedure, during which participants processed language stimuli. **A.c** illustrates the standard brain encoding framework. **B** Brain scores across stories computed from representations at different model layers. Based on these results, we select the middle one-third of layers to compute whole-story brain scores, shown in **C**. The 20th layer, which consistently achieves the highest scores, is further selected for subject-wise analysis. Statistical significance is assessed using the Wilcoxon rank-sum test.

3.1 Methods

This section describes how we obtain and use joint vision-language representations to predict neural responses to language stimuli. While participants receive language-only stimuli during fMRI recording, we first apply visual augmentation to the text, then extract joint vision-language representations, and finally use these representations in an encoding model to predict neural activity.

Step 1: Visual Augmentation. As discussed above, most paired image-text datasets are designed for studying neural responses to visual stimuli, where text typically appears as image descriptions. In contrast, language neuroscience experiments often use only continuous natural language stimuli, which are not directly visualizable. To address this challenge, we propose a visual augmentation procedure for extracting joint vision-language representations from text.

Given a continuous narrative text S , we first segment it into a sequence of short clauses using punctuation that introduces natural pauses (e.g., commas and periods):

$$S = \{S_1, S_2, \dots, S_n\}. \quad (1)$$

We then use an image generation model G to generate a visual scene for each clause. Image gen-

eration models typically take as input a content description, contextual information, and optional constraints on style or resolution. For each clause S_t ($1 \leq t \leq n$), we use S_t as the content description and the preceding clauses $\{S_1, \dots, S_{t-1}\}$ as contextual input. The generated image is given by:

$$I_t = G(S_t; \{S_1, \dots, S_{t-1}\}). \quad (2)$$

This procedure yields a set of paired text-image stimuli:

$$S' = \{(S_1, I_1), (S_2, I_2), \dots, (S_n, I_n)\}. \quad (3)$$

This design accounts for the fact that some clauses in narrative text lack explicit visual content. Conditioning on preceding context allows such clauses to be visualized in a plausible manner. For example, when characters exchange greetings such as “Hello,” the generated image depicts two people interacting rather than a literal visualization of the word itself.

Step 2: Joint Representation Extraction. We next extract joint representations from the paired text-image stimuli S' . We use a MLLM, which typically consists of a vision encoder for visual feature extraction and a multi-layer backbone that jointly processes visual features and textual embeddings.

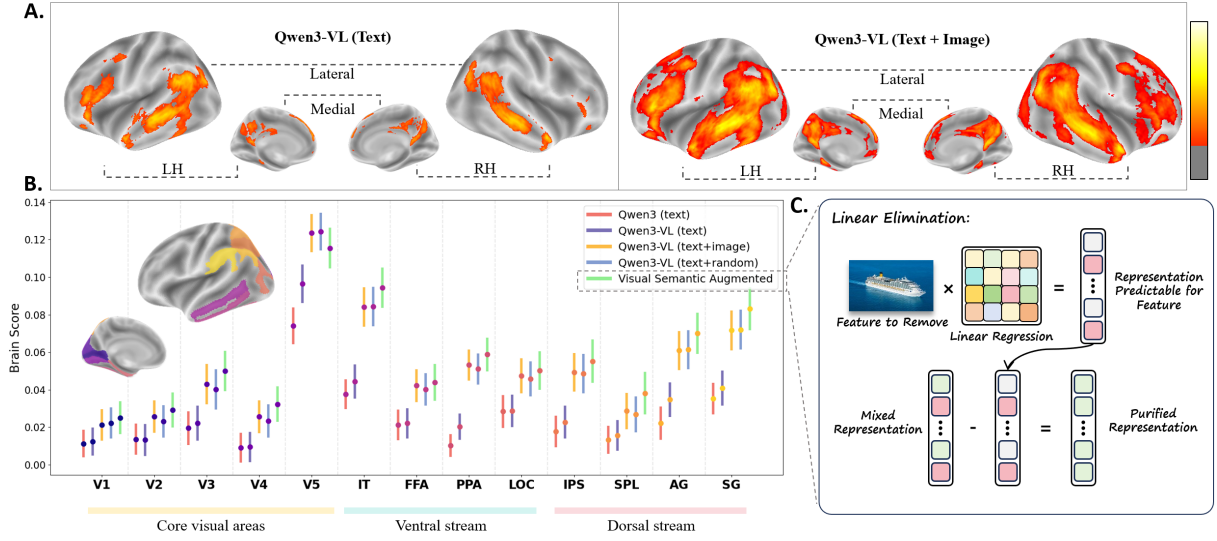


Figure 3: (A) shows voxel-wise brain scores obtained from language-only and vision-language joint representations extracted from the 20th layer of Qwen3-VL. Lighter colors indicate higher predictive accuracy for language-evoked neural responses. (B) reports region-of-interest (ROI)-level brain scores, with error bars denoting 95% confidence intervals. Among the compared conditions, the visual semantic augmented representation is constructed using the linear removal method, whose overall procedure is illustrated in (C).

Formally, each S_t is represented as a sequence of words:

$$S_t = \{w_1^t, w_2^t, \dots, w_m^t\}, \quad (4)$$

where w_k^t denotes the k -th word in the clause. The word sequence, together with its corresponding image I_t and preceding textual context, is fed into the MLLM M .

We extract representations from the hidden layers of the model. Specifically, the representation of the k -th word is defined as:

$$r_k^t = M(w_1^t, w_2^t, \dots, w_k^t; I_t), \quad (5)$$

where r_k^t denotes the hidden-state representation associated with word w_k^t .

This representation jointly encodes information from the current word, preceding linguistic context, and the visual modality. We therefore treat r_k^t as a word-level vision-language joint representation.

Step 3: Brain Encoding for Predicting Neural Responses. Given the extracted representations, we aim to predict brain activity recorded as functional magnetic resonance imaging (fMRI) signals elicited by language-only stimuli using a standard brain encoding framework.

fMRI signals are sampled at discrete repetition times (TRs), during which participants typically process multiple words. We construct a

TR-level stimulus representation by summing the representations of all words presented within the same TR. Let r_k^t denote the joint vision-language representation of word w_k^t extracted as described above. For example, if a TR contains the words $\{w_{m-1}^t, w_m^t, w_1^{t+1}\}$, the corresponding stimulus representation is defined as:

$$\mathcal{R} = r_{m-1}^t + r_m^t + r_1^{t+1}. \quad (6)$$

Brain encoding maps the representation $\mathcal{R} \in \mathbb{R}^d$ of a stimulus sequence $\mathcal{S}$ to the elicited brain responses $\mathcal{A} \in \mathbb{R}^v$, where v denotes the number of voxels (Figure 2A.c). We model this relationship using a linear encoding model:

$$\hat{\mathcal{A}} = \mathbf{w} \cdot \text{FIR}(\mathcal{R}) + \mathbf{b}, \quad (7)$$

where $\mathbf{w} \in \mathbb{R}^{v \times d}$ is the regression weight matrix and $\mathbf{b}$ is the bias term. Model parameters are estimated by minimizing an l_2 -regularized least-squares objective:

$$\arg \min_{\mathbf{w}, \mathbf{b}} \sum_{i=1}^N \left(\mathcal{A}_i - \hat{\mathcal{A}}_i \right)^2 + \lambda \|\mathbf{w}\|_2^2. \quad (8)$$

To account for the temporal delay of the BOLD response, we apply a finite impulse response (FIR) model to $\mathcal{R}$, following prior work (Shain et al., 2020; Deniz et al., 2019). Representations from

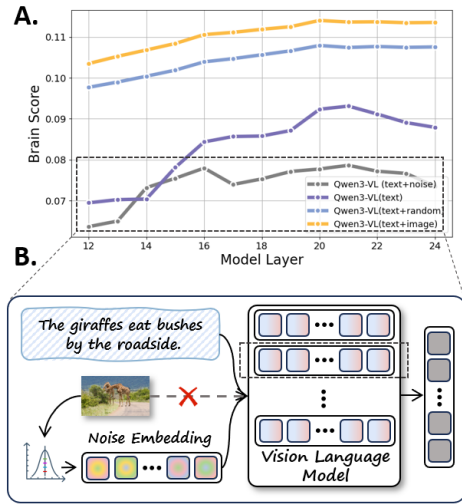


Figure 4: Brain encoding performance using representations extracted by combining text with random noise vectors, compared to text-only and text-image conditions. The inclusion of pure mathematical noise degrades encoding efficacy, ruling out high-dimensional feature expansion artifacts.

multiple TRs are concatenated using a 6-second (5-TR) temporal window. Model performance is evaluated by the Pearson correlation between predicted and measured fMRI responses for each voxel, referred to as the **brain score**.

Furthermore, to assess the theoretical upper bound of our linear encoding models, we computed the inter-subject noise ceiling. Detailed methodology and whole-brain noise ceiling maps are provided in Appendix E.

4 Experimental Settings

In this section, we apply the proposed method to neural datasets to test whether vision-language joint representations better predict neural responses to language-only stimuli. We describe the datasets, models, and the control conditions for comparison.

Datasets. We analyzed the publicly available Narratives dataset (Nastase et al., 2021), which contains fMRI recordings from 345 subjects listening to 27 English stories of ~ 7 –56 minutes each (totaling ~ 4.6 hours and $\sim 43,000$ words). We used preprocessed fMRI without spatial smoothing and excluded noisy or incomplete story–subject pairs, yielding 618 pairs and ~ 4 hours of audio. Since participants received language stimuli exclusively via the auditory modality, potential effects of visual word form processing on visual pathways were

excluded (Constable et al., 2004).

Multimodal Large Language Models. We evaluate three representative families of open-source MLLMs, all built on high-capacity large language models (LLMs), which are also included as unimodal baselines. **Qwen.** The multimodal model is Qwen3-VL, with Qwen3 as its unimodal counterpart and shared language backbone. **LLaVA.** We use the latest LLaVA-OneVision model as the multimodal variant, built on the Qwen2 language backbone. **LLaMA.** The multimodal model is LLaMA-3.2-Vision, which uses LLaMA-3.1 as its language backbone. Across all families, we use relatively small variants with approximately 7-12B parameters. All models are initialized from publicly available pretrained checkpoints obtained from Hugging Face. In the main text, we report results from the Qwen family as a representative example, with results from the other two families provided in the appendix H. The main findings are consistent across all model families.

Regression Model. We train brain encoding models using voxel-wise Ridge regression (RidgeCV) implemented in the himalaya package. The regularization parameter α is selected from a logarithmically spaced grid $\text{logspace}(-1, 8, 10)$. Model performance is evaluated using 20-fold cross-validation, with the Pearson correlation between predicted and measured fMRI responses computed for each voxel as the brain score.

Control Conditions. To systematically assess the contribution of vision-language joint representations to predicting neural responses to language, we define four experimental conditions (Figure 2A.a). (1) **LLM (Text):** unimodal LLM with language-only input. (2) **MLLM (Text):** MLLM receives only language stimuli. (3) **MLLM (Text + Image) (Main Condition):** our proposed setting, in which a MLLM processes visually augmented text-image pairs to extract joint representations. (4) **MLLM (Text + Random):** language stimuli are paired with semantically unrelated images and jointly fed into the multimodal model, testing whether performance gains depend on image semantics rather than the mere presence of visual input. To ensure the validity of this control condition, we quantitatively verified that the random images are strictly semantically decoupled from the linguistic stimuli. A detailed semantic alignment evaluation, confirming a near-zero semantic overlap for random im-

ages compared to the robust alignment of matched images, is provided in Appendix B.

Because image generation is stochastic, we generate five independent image sets for each text stimulus, extract representations separately, and average the resulting vectors.

5 Preliminary Results and Findings

In this section, we report the main experimental results and summarize three primary empirical findings based directly on observations, without introducing new hypotheses.

Figure 2B shows brain scores across hidden layers for all story stimuli. Consistent with prior brain encoding studies, performance increases in earlier layers and declines later. Consequently, we focus on the middle third of the layers and concatenate all stories for joint prediction (Figure 2C). Layer 20 yields consistently high performance, which we use to predict individual neural responses (Figure 2D), assessing relative differences via the Wilcoxon rank-sum test. This yields the following findings.

Finding 1: Multimodal models outperform unimodal models on text processing. Even without visual input, MLLMs improve brain scores by nearly 20% over unimodal LMs when predicting language-evoked responses (Figure 2C). However, this gain is modest compared to the improvements from explicitly incorporating visual information.

Finding 2: Multimodal pathway activation yields further gains. Supplying visual input to MLLMs triggers multimodal processing pathways, further boosting brain encoding performance. This improvement surpasses the text-only MLLM baseline (Figure 2C) regardless of image-text semantic alignment. Specifically, matched image-text representations increase brain scores by nearly 25% over text-only inputs, while random images still yield an approximately 18% gain. This echoes prior findings that language alignment is not strictly necessary to improve neural predictions for visual stimuli (Wang et al., 2023; Conwell et al., 2022).

Finding 3: Visual information injection selectively enhances predictions in visual regions. At the brain-region level (Figure 3B), score increases from added visual information are concentrated in visual cortices. The largest gains—nearly doubling original values—occur in mid- and high-level visual areas, spanning dorsal (posterior temporal cortex) and ventral streams (intraparietal sul-

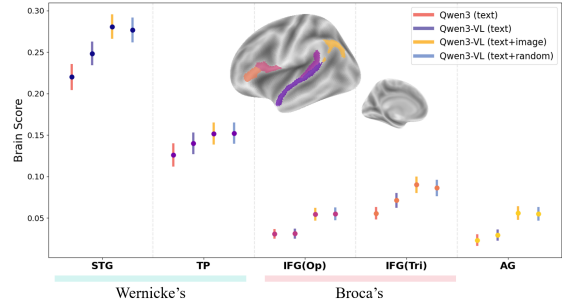


Figure 5: Brain scores of language-related regions under different experimental conditions. Notably, STG and AG are associated with both language processing and visual-semantic functions.

cus), including V5, IT, PPA, and Inferior Parietal Lobule (AG and SG). Canonical language regions (Figure 5) maintain relatively high baseline scores due to the stimuli’s linguistic nature and MLLM training, thus showing smaller relative gains. However, regions contributing to multimodal processing, such as STG and AG, exhibit more pronounced improvements. Overall, joint representations enhance predictions across all regions, but effects peak in visual areas.

6 In-Depth Analysis of the Findings

Building on our finding that joint vision-language representations from MLLMs improve neural predictions primarily in visually related regions, we investigate the underlying mechanisms. We first outline foundational assumptions, then examine this from two complementary analytical perspectives (APs): (AP1) the neural mechanisms underlying human language comprehension, and (AP2) the representational properties of MLLMs. For each, we introduce supplementary experiments and inferences, ultimately integrating these insights into a unified conclusion.

6.1 Consensus Assumptions

We rely on widely accepted neuroscientific and machine learning premises as conceptual foundations, acknowledging that their precise detailed mechanisms remain active research areas.

Assumption 1: The human brain contains specialized networks for visual processing and cross-modal integration. We adopt a framework dividing the visual system into three interconnected components (Braddick and Atkinson, 2011): core visual areas (early processing), the ventral stream (object recognition), and the dorsal stream (spa-

tial processing, scene understanding, and cross-modal integration). These functional distinctions—including overlapping regions like V4 and V5—are essential for interpreting our subsequent region-specific analyses.

Assumption 2: MLLMs can incorporate visual information into textual representations. While seemingly intuitive, MLLMs can exhibit hallucination behaviors, over-relying on linguistic priors and underutilizing visual inputs. Therefore, we cannot guarantee that visual content is always effectively encoded. We treat this assumption as plausible but not assured, remaining open to alternative interpretations when analyzing the sources of performance gains.

6.2 Multimodal Neural Mechanisms in the Human Brain (AP1)

Joint representations improve neural predictions for language, suggesting visual semantic recruitment during language processing, aligning with neuroscience literature (Bedny et al., 2011; Seydell-Greenwald et al., 2023). However, determining if these gains stem specifically from visual semantics is challenging. While matched images yield the highest brain scores in mid- and high-level visual regions, semantically mismatched images also provide non-negligible improvements. To clarify this, we analyze the joint representations to address two confounders: (1) whether gains are mathematical artifacts of adding high-dimensional noise, and (2) how to isolate purely semantic visual contributions.

Additional experiments. We introduce two supplementary analyses to rigorously evaluate these gains. First, to rule out mathematical artifacts from high-dimensional feature expansion or implicit regularization, we introduce a *pure mathematical noise* control. We replace image inputs with random noise vectors—sampled from the actual image embedding distribution—combined with text to extract MLLM joint embeddings (Figure 4B). Second, to isolate semantically relevant visual inputs, we use *linear elimination* (Oota et al., 2023) (Figure 3C) to suppress gains from mismatched (random) images. If a representation $\mathcal{R}$ linearly encodes a feature $\mathcal{F}$ such that $\mathcal{R}_{\mathcal{F}} = \mathcal{L}(\mathcal{F})$ via a linear model $\mathcal{L}$, linear elimination ($\ominus$) computes the residual:

$$\mathcal{R} \ominus \mathcal{F} = \mathcal{R} - \mathcal{L}(\mathcal{F}), \quad (9)$$

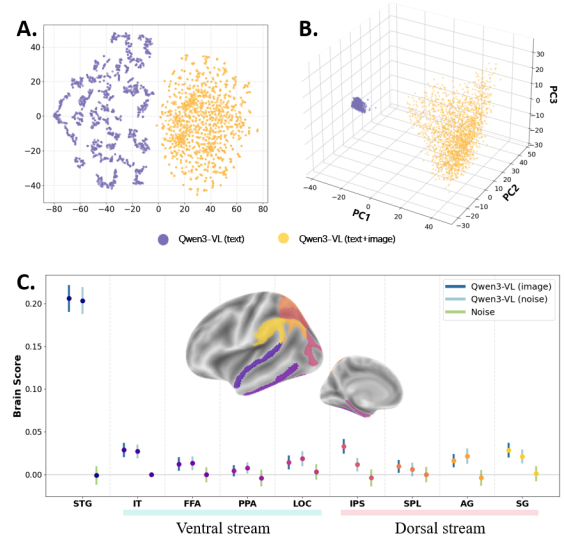


Figure 6: **A & B** detail the representational geometry of language-only and joint representations from Qwen3-VL via t-SNE embeddings (**A**) and principal component distributions (**B**). **C** reports brain scores using images versus random noise vectors to predict language-evoked neural responses.

removing linearly decodable information about $\mathcal{F}$. Because visual inputs may interact with text internally, we avoid directly removing mismatched image features. Instead, the purified joint representation is defined as:

$$\mathcal{R}_{\text{joint}} = \mathcal{R}_{\text{text+image}} \ominus \left(\mathcal{R}_{\text{text+random}} \ominus \mathcal{R}_{\text{text}} \right), \quad (10)$$

where $\mathcal{R}_{\text{text+random}} \ominus \mathcal{R}_{\text{text}}$ isolates non-semantic visual information, removing it from $\mathcal{R}_{\text{text+image}}$ to suppress non-semantic visual contributions. Furthermore, to address potential limitations of linear approximation in isolating deep non-linear visual features, we complement this approach with Representational Similarity Analysis (RSA). RSA directly evaluates the geometric alignment between the model’s internal representational space and the brain’s response space without relying on linear regression. Detailed methodology and results for RSA are provided in Appendix F.

Inferences. The pure noise condition (Figure 4) reveals that adding random noise vectors substantially degrades brain encoding performance compared to the text-only baseline. This rules out generic representational effects (e.g., regularization from extra tokens), confirming that score improvements rely on structured visual features aligning with human visual pathways. Consequently, the improved predictions in visual cortices from ac-

Table 1: Geometric and information-theoretic statistics of representations under different settings.

Setting	Geom.		Info.
	L2 Norm	Var	95% PC
LM (text)	4690	2.54	431
VLM (text)	4486	2.32	239
VLM (text + rand)	4796	2.64	1033
VLM (text + img)	4764	2.71	1143

tual images (Figure 3B) suggest the brain engages visual pathways during language processing, signals which joint representations effectively capture. Furthermore, applying linear elimination to remove non-semantic visual information further increases brain scores across nearly all mid- and high-level visual regions (both ventral and dorsal streams, except V5). This indicates visual pathway activation is specifically tied to semantically relevant visual information. Because these regions support object recognition and scene understanding, this implies brain generates semantic-specific visual associations during language comprehension.

6.3 Representational Mechanisms in MLLMs (AP2)

To determine if MLLMs exhibit brain-like representational mechanisms, we analyze how modality-specific and joint representations interact within their shared backbone.

Additional experiments. We introduce two conditions to probe representational dynamics: an *image-only* setting (text-matched images) to isolate visual representations, and a *noise input* setting (random vectors matching image embedding distributions) to capture uninformative input evolution. To characterize representational geometry, we compute L2 norm, variance, and the number of principal components (PCs) explaining 95% of variance (measuring information density).

Inferences. Despite lacking semantic content, random vectors fed into MLLMs yield representations structurally similar to human cross-modal integration regions, notably the STG (McGurk and MacDonald, 1976). This implies MLLMs possess an inherently multimodal representational geometry. Adding actual images further aligns representations with the dorsal visual pathway, while combined text-image inputs extend this alignment to both dorsal and ventral high-level visual areas. Structurally, visual information slightly increases

L2 norm and variance but more than triples the PCs required to explain 95% of the variance (Table 1). Unimodal language representations typically exhibit representation anisotropy (Ethayarajh, 2019), clustering tightly in t-SNE space with variance concentrated in few dimensions (Timkey and van Schijndel, 2021). Visual input mitigates this redundancy, dispersing representations more evenly and reducing the model’s reliance on fine-grained language features. From a brain encoding perspective, this dispersed, enriched geometric space directly facilitates linear regression, thereby improving predictive performance.

6.4 Mechanism Summary

Based on the above analyses, we summarize the mechanism by which vision–language joint representations enhance brain encoding. First, linguistic processing in the human brain co-activates visual pathways, generating signals that can be predicted from visual information. Concurrently, MLLMs encode visual inputs into representations that are linearly aligned with human brain representational spaces, likely reflecting similarities in intrinsic geometry between the model and brain regions involved in visual–language integration. In particular, visual information is mapped into a space corresponding to visual cortical areas, especially dorsal stream regions. Combining visual and linguistic input further strengthens this alignment, potentially by reducing representation anisotropy induced by visual information.

7 Discussion

This work demonstrates that injecting visual information into language representations significantly improves the prediction of language-evoked brain activity. These gains stem from the engagement of visual pathways during comprehension and the alignment between MLLM latent spaces and visual cortices. Crucially, our auditory narrative paradigm eliminates bottom-up visual interference, ensuring the observed predictive gains reflect pure top-down visual-semantic associations. Beyond cognitive neuroscience, our findings offer actionable insights for NLP: text-only LLMs could benefit from “visual-spatial hallucination” latent layers to improve physical grounding, enhancing multimodal pathway utilization during unimodal tasks.

Limitations

While our findings provide robust evidence for the multimodal nature of language processing and model representations, several limitations remain. First, although MLLMs offer greater representational capacity, their internal mechanisms are highly non-linear and difficult to interpret. We observed that injecting mismatched images into model inputs also yields generic representational expansions that slightly increase brain scores. While our linear elimination method theoretically isolates the semantic visual contributions, it serves as a first-order approximation. Deep non-linear attention interactions between text and image tokens may leave residual text information in the purified representations. Although we mitigate this by employing Representational Similarity Analysis (RSA) to corroborate geometric alignment without relying on linear assumptions, future work should explore advanced causal ablation techniques to fully disentangle these non-linear interactions at a mechanistic level. Second, our empirical analyses are grounded in the Narratives dataset, which comprises exclusively English audio stories. Although visual imagery and multimodal grounding are widely considered fundamental mechanisms of human cognition that are largely language-independent, the specific neural dynamics may vary. Future research should extend this multimodal brain encoding framework to reading-based fMRI datasets and diverse, non-English languages to fully validate the universality of these brain-model alignment phenomena.

Acknowledgements

This research is supported by the National Natural Science Foundation of China (No.62476127,62306140), the Natural Science Foundation of Jiangsu Province (No.BK20242039), the Fundamental Research Funds for the Central Universities(No.NXD2026006), Disciplinary Basic Research Project Funding, (No.ILF26023), the Research Fund (No.PO250624101698), the Scientific Research Starting Foundation of Nanjing University of Aeronautics and Astronautics (No.YQR21022), and the High Performance Computing Platform of Nanjing University of Aeronautics and Astronautics.

References

- Marina Bedny, Alvaro Pascual-Leone, David Dodell-Feder, Evelina Fedorenko, and Rebecca Saxe. 2011. Language processing in the occipital cortex of congenitally blind adults. *Proceedings of the National Academy of Sciences*, 108(11):4429–4434.
- Oliver Braddick and Janette Atkinson. 2011. Development of human visual function. *Vision research*, 51(13):1588–1609.
- Charlotte Caucheteux and Jean-Rémi King. 2022. Brains and algorithms partially converge in natural language processing. *Communications biology*, 5(1):134.
- R Todd Constable, Kenneth R Pugh, Ella Berroya, W Einar Mencl, Michael Westerveld, Weijia Ni, and Donald Shankweiler. 2004. Sentence complexity and input modality effects in sentence comprehension: an fmri study. *NeuroImage*, 22(1):11–21.
- Colin Conwell, Jacob S Prince, Kendrick N Kay, George A Alvarez, and Talia Konkle. 2022. What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *BioRxiv*, pages 2022–03.
- Fatma Deniz, Anwar O Nunez-Elizalde, Alexander G Huth, and Jack L Gallant. 2019. The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality. *Journal of Neuroscience*, 39(39):7722–7736.
- Barry J Devereux, Alex Clarke, Andreas Marouchos, and Lorraine K Tyler. 2013. Representational similarity analysis reveals commonalities and differences in the semantic processing of words and objects. *Journal of Neuroscience*, 33(48):18906–18916.
- Adrien Doerig, Tim C Kietzmann, Emily Allen, Yi-han Wu, Thomas Naselaris, Kendrick Kay, and Ian Charest. 2022. Semantic scene descriptions as an objective of human vision. *arXiv preprint arXiv:2209.11737*, 10.
- Dota Tianai Dong and Mariya Toneva. 2023. Vision-language integration in multimodal video transformers (partially) aligns with the brain. *arXiv preprint arXiv:2311.07766*.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.
- Michael Eickensberg, Alexandre Gramfort, Gaël Varoquaux, and Bertrand Thirion. 2017. Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, 152:184–194.
- Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In

- Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.
- Shailee Jain and Alexander Huth. 2018. Incorporating context into language encoding models for fmri. *Advances in neural information processing systems*, 31.
- Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. 2008. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249.
- Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.
- Harry McGurk and John MacDonald. 1976. Hearing lips and seeing voices. *Nature*, 264(5588):746–748.
- Juliette Millet, Charlotte Caucheteux, Yves Boubenec, Alexandre Gramfort, Ewan Dunbar, Christophe Pallier, Jean-Remi King, et al. 2022. Toward a realistic model of speech processing in the brain with self-supervised learning. *Advances in Neural Information Processing Systems*, 35:33428–33443.
- Bence Nanay. 2018. Multimodal mental imagery. *Cortex*, 105:125–134.
- Samuel A. Nastase, Yun-Fei Liu, Hanna Hillman, Asieh Zadbood, Liat Hasenfratz, Neggin Keshavarzian, Janice Chen, Christopher J. Honey, Yaara Yeshurun, Mor Regev, Mai Nguyen, Claire H. C. Chang, Christopher Baldassano, Olga Lositsky, Erez Simony, Michael A. Chow, Yuan Chang Leong, Paula P. Brooks, Emily Micciche, and 6 others. 2021. [Narratives: fmri data for evaluating models of naturalistic language comprehension](#). *bioRxiv*.
- Subba Reddy Oota, Jashn Arora, Vijay Rowtula, Manish Gupta, and Raju S Bapi. 2022. Visio-linguistic brain encoding. *arXiv preprint arXiv:2204.08261*.
- Subba Reddy Oota, Khushbu Pahwa, Mounika Marreddy, Maneesh Singh, Manish Gupta, and Bapi S Raju. 2025. Multi-modal brain encoding models for multi-modal stimuli. *arXiv preprint arXiv:2505.20027*.
- SubbaReddy Oota, Manish Gupta, and Mariya Toneva. 2023. Joint processing of linguistic properties in brains and language models. *Advances in Neural Information Processing Systems*, 36:18001–18014.
- Sara F Popham, Alexander G Huth, Natalia Y Bilenko, Fatma Deniz, James S Gao, Anwar O Nunez-Elizalde, and Jack L Gallant. 2021. Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nature neuroscience*, 24(11):1628–1636.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.
- Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. 2018. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, page 407007.
- Anna Seydell-Greenwald, Xiaoying Wang, Elissa L Newport, Yanchao Bi, and Ella Striem-Amit. 2023. Spoken language processing activates the primary visual cortex. *PLoS One*, 18(8):e0289671.
- Cory Shain, Idan Asher Blank, Marten van Schijndel, William Schuler, and Evelina Fedorenko. 2020. fmri reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, 138:107307.
- Charles Spence and Ophelia Deroy. 2012. Crossmodal mental imagery. *Multisensory imagery*, pages 157–183.
- Jerry Tang, Meng Du, Vy Vo, Vasudev Lal, and Alexander Huth. 2023. Brain encoding models based on multimodal transformers can transfer across language and vision. *Advances in Neural Information Processing Systems*, 36:29654–29666.
- Bertrand Thirion, Edouard Duchesnay, Edward Hubbard, Jessica Dubois, Jean-Baptiste Poline, Denis LeBihan, and Stanislas Dehaene. 2006. Inverse retinotopy: inferring the visual content of images from brain activation patterns. *Neuroimage*, 33(4):1104–1116.
- William Timkey and Marten van Schijndel. 2021. [All bark and no bite: Rogue dimensions in transformer language models obscure representational quality](#). *CoRR*, abs/2109.04404.
- Mariya Toneva and Leila Wehbe. 2019. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Advances in neural information processing systems*, 32.
- Greta Tuckute, Jenelle Feather, Dana Boebinger, and Josh H McDermott. 2023. Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions. *Plos Biology*, 21(12):e3002366.

- Aditya R Vaidya, Shailee Jain, and Alexander G Huth. 2022. Self-supervised models of audio effectively explain human cortical responses to speech. *arXiv preprint arXiv:2205.14252*.
- Aria Y Wang, Kendrick Kay, Thomas Naselaris, Michael J Tarr, and Leila Wehbe. 2023. Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nature Machine Intelligence*, 5(12):1415–1426.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. 2014. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS one*, 9(11):e112575.
- Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624.

A Image Generation

Image Generation Details. We used the **FLUX.1-dev** model for image generation. We also experimented with **Stable-Diffusion-3.5-Large**; however, visual inspection revealed that its generated images were of lower quality than those produced by FLUX, particularly in cases without explicit scene descriptions (e.g., dialogues). In such settings, Stable Diffusion often produced unnatural or implausible images. Therefore, all experiments reported in this paper exclusively use FLUX.1-dev.

The FLUX model was downloaded from Hugging Face with pretrained weights. To accelerate generation, we employed **4-bit quantization** with the **NF4** format. Manual evaluation indicated that this quantization strategy had negligible impact on generation quality.

Prompt Design and Generation Procedure. For each short sentence extracted from the linguistic stimulus, we constructed prompts that incorporate both the current sentence and its preceding context. Specifically, let `seq` denote the current short sentence, which directly describes the scene to be generated, and let `context` denote the preceding sentences that provide narrative background. The core generation code is shown below:

```
prompt = f"Scene described in this sentence: "
        f"{seq}\n{style_prompt}"

prompt_2 = (
    f"Context: {context}\n"
    f"Scene described in this sentence: {seq}\n"
    "The image should reflect the described "
    "situation naturally, no words or text."
)

image = pipe(
    prompt,
    prompt_2=prompt_2,
    negative_prompt=(
        "illustration, painting, drawing, sketch, "
        "digital painting, anime, cartoon, "
        "3D render, low quality, blurry, "
        "text artifacts, words or text"
    ),
    height=HEIGHT,
    width=WIDTH,
    guidance_scale=GUIDANCE_SCALE,
    num_inference_steps=STEPS,
    max_sequence_length=512,
    generator=generator
).images[0]
```

Here, `seq` specifies the target content of the image, while `context` provides additional narrative information to ensure coherent and natural scene generation, especially for sentences that are not directly visualizable in isolation. The negative prompt was designed to suppress non-photorealistic styles, textual artifacts, and low-quality generations.

Qualitative Results. We present a subset of the generated images to illustrate the effectiveness of the proposed visual augmentation procedure (Figure 7).

B Semantic Validity of Visual Augmentation

To address potential concerns regarding the semantic independence of the random-image control and to quantitatively validate our image generation pipeline, we evaluate the semantic alignment between the linguistic stimuli and the visual inputs. Specifically, we measure the semantic similarity under both the matched-image and random-image conditions using two complementary approaches.

First, we compute the direct cosine similarity between the embeddings of the original text stimuli and the corresponding image embeddings. Second, to provide a more interpretable text-to-text comparison, we perform a caption recovery test. We use a large multimodal model to generate captions for the visual inputs and then calculate the text embedding cosine similarity between these generated captions and the original sentences.

The quantitative results are summarized in Table 2. The matched images exhibit an average cosine similarity of 0.54 with the text embeddings, whereas the random images yield a near-zero similarity of 0.08. Furthermore, in the caption recovery test, the matched images achieve a similarity score of 0.61. Considering that a significant portion of the Narratives dataset consists of conversational or non-descriptive language, this score indicates a robust preservation of core narrative semantics. In contrast, the random images yield a caption-to-text similarity of only 0.12.

These results empirically confirm that our visually augmented images successfully capture the semantic content of the text. More importantly, they verify that the random images are fundamentally decoupled from the narrative semantics, ensuring that the performance gains observed in the “Text + Random Image” condition are not driven by coincidental semantic overlap.

C Control for Potential Temporal Leakage

A potential concern in our visual augmentation pipeline is that the image associated with a clause is generated using the complete clause, which may indirectly expose information from words occur-



Figure 7: A subset of the generated images.

Table 2: Quantitative evaluation of semantic alignment. Image–Text denotes the average cosine similarity between the original text and image embeddings, while Caption–Text denotes the cosine similarity between the original text and the captions generated from the images.

Condition	Similarity	
	Image–Text	Caption–Text
Text + Matched Image	0.54	0.61
Text + Random Image	0.08	0.12

ring later within the same clause. Since previous work has shown that incorporating future linguistic context can improve neural response prediction, we conducted an additional control experiment to estimate the magnitude of this potential effect.

Experimental Setup. We explicitly introduced future linguistic information into the text-only condition by incorporating embeddings from words occurring after the current prediction position. All other components of the encoding pipeline were kept unchanged. This control provides an empirical

Table 3: Control experiment for potential temporal leakage. Explicitly adding future-word information produces only a small improvement compared with visual augmentation.

Condition	Brain Score
Text Only	0.078
Text + Future Words	0.082
Text + Matched Image	0.119

estimate of how much prediction performance can be improved solely by access to future linguistic information, without introducing any visual input.

Results. The results are summarized in Table 3. The brain score of the text-only condition is 0.078. Explicitly incorporating future-word embeddings increases the score to 0.082, corresponding to an absolute improvement of only 0.004. In contrast, the matched-image condition achieves a brain score of 0.119, yielding an improvement of 0.041 over the original text-only condition.

We further examined the spatial distribution of the improvement induced by future-word information. The gain was mainly concentrated in tem-

Table 4: Generated-caption control on the *Pieman* story. Replacing visual augmentation with additional textual descriptions does not reproduce the improvement obtained from visual inputs.

Condition	Brain Score
Original Text	0.121
Text + Generated Caption	0.101

poral regions associated with language processing, whereas the improvements induced by matched visual augmentation extend to middle- and high-level visual and multimodal regions. These results suggest that access to future linguistic information can provide a small prediction advantage, but its magnitude and spatial distribution are substantially different from the effects observed under visual augmentation.

D Textual Semantic Expansion Control

To examine whether the improvement from visual augmentation can be reproduced by simply providing additional semantically related text, we conducted a preliminary generated-caption control experiment on the *Pieman* story.

Experimental Setup. For each generated image, we first used a multimodal model to recover a textual caption describing the visual content. The generated caption was then appended to the corresponding textual input, after which the resulting sequence was encoded using the same representation extraction and fMRI prediction pipeline. This condition preserves the semantic content introduced by the generated image while replacing the visual pixels with a textual description.

Results. As shown in Table 4, adding the generated captions does not improve neural prediction performance. The original text condition achieves a brain score of 0.121, whereas appending the generated image captions decreases the score to 0.101, corresponding to a reduction of approximately 0.020.

Although this experiment is conducted on a single story and should therefore be interpreted as a preliminary control, the result indicates that the gain from visual augmentation is unlikely to be explained by a trivial increase in the amount of semantically related textual information.

E Noise Ceiling Estimation

To evaluate the theoretical upper bound of the linear encoding models, we computed the noise ceiling for the fMRI data. The noise ceiling estimates the maximum possible predictive performance given the inherent noise and trial-to-trial variability in the BOLD signal.

Methodology. We calculated the noise ceiling using a standard inter-subject correlation approach. Specifically, for each subject, we predicted their voxel-wise fMRI responses using the averaged fMRI responses of all other subjects who listened to the exact same narrative stimuli. The resulting Pearson correlation coefficient between the averaged actual responses of the group and the target subject serves as the noise ceiling for that subject. This value was then averaged across all subjects to yield the final group-level noise ceiling map.

Results and Inferences. Figure 8 presents the noise ceiling estimates across the cortex and within predefined ROIs. As shown in the whole-brain maps (Figure 8A), the theoretical upper bound is highest in the auditory cortex and canonical language networks, reflecting a high degree of inter-subject consistency during auditory narrative comprehension.

Within the visual processing pathways (Figure 8B), the noise ceiling varies significantly. The highest upper bounds are observed in regions associated with cross-modal integration and high-level visual processing, notably the Angular Gyrus (AG, ≈ 0.37), Middle Temporal visual area (V5, ≈ 0.33), and Supramarginal Gyrus (SG, ≈ 0.29). In contrast, core visual areas (V1–V4) exhibit relatively lower noise ceilings (ranging from 0.12 to 0.17).

This distribution aligns with our experimental paradigm: since the stimuli are language-only audio narratives without direct low-level visual input, early visual cortices display lower stimulus-driven inter-subject consistency. Importantly, the robust noise ceilings observed in mid- and high-level visual regions (e.g., AG and V5) validate our main encoding results in Section 5. It confirms that the substantial performance gains achieved by multimodal representations in these areas correspond to reliable, cross-subject semantic variance rather than the fitting of random noise.

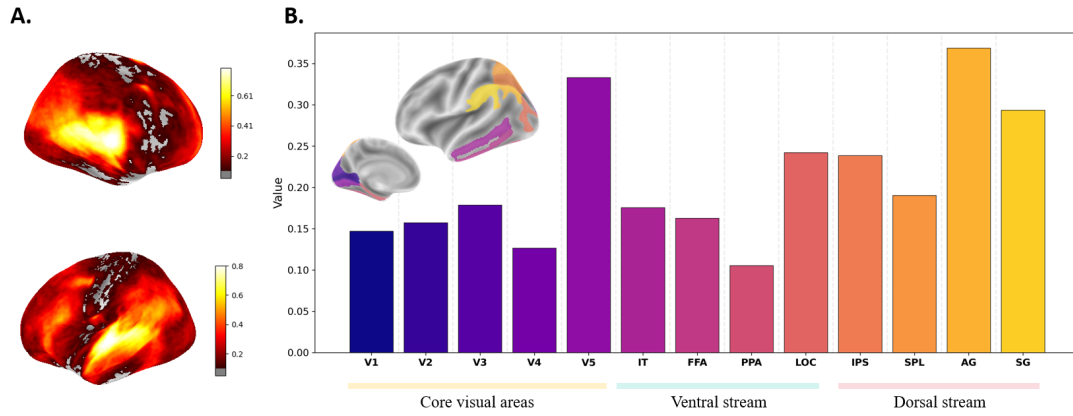


Figure 8: **A.** Whole-brain visualization of the estimated noise ceilings, indicating the theoretical upper bound for linear encoding models across the cortex. Lighter colors indicate higher inter-subject consistency. **B.** Region-of-interest (ROI) level noise ceilings along the visual processing pathways.

F Representational Similarity Analysis (RSA)

Motivation. In Section 6.2, we utilized a linear elimination approach to isolate the contribution of semantically relevant visual inputs. However, given that MLLMs exhibit highly complex, non-linear interactions between visual and textual information, linear subtraction may serve only as a first-order approximation. To complement the encoding analysis and directly examine the geometry of the learned representations, we conduct a Representational Similarity Analysis (RSA) (Kriegeskorte et al., 2008). RSA compares the relational structure of model representations with that of neural responses without fitting an explicit mapping between the two spaces.

Methodology. For each experimental condition, we computed a Representational Dissimilarity Matrix (RDM) from the stimulus representations extracted from the MLLM. The dissimilarity between two representations was defined as one minus their Pearson correlation coefficient. Similarly, we computed neural RDMs from the fMRI BOLD responses within the selected visually implicated ROIs. The RSA score was then defined as the Spearman rank correlation between the upper triangular elements of the model RDM and the corresponding neural RDM.

Layer-wise Results. As shown in Figure 9, both vision-language joint representations (*Text + Matched Image* and *Text + Random Image*) achieve substantially higher RSA scores than the language-only baseline across the analyzed layers. This indicates that introducing visual inputs changes the

global representational geometry of the MLLM in a direction that is more similar to the geometry of neural responses.

Interestingly, the *Text + Random Image* condition achieves slightly higher RSA scores than the *Text + Matched Image* condition in several layers. This suggests that part of the RSA improvement reflects a general geometry-reshaping effect induced by visual input rather than semantic alignment alone. We therefore conducted additional controls to separate this generic visual effect from the contribution of semantically matched visual information.

Fine-grained Control Analysis at Layer 20. Table 5 summarizes the RSA scores of several control conditions at layer 20. Adding an arbitrary noise vector to the text representation decreases the RSA score from 0.278 to 0.254, indicating that the improvement observed with visual input cannot be explained by simply injecting additional high-dimensional noise.

In contrast, both matched and random images substantially increase the RSA score, reaching 0.343 and 0.360, respectively. Importantly, after linearly removing the component associated with the random-image condition, the purified joint representation still achieves an RSA score of 0.345, which remains substantially higher than the text-only baseline.

These results indicate that generic visual input can reshape the global representation geometry and thereby improve RSA alignment. At the same time, the strong RSA score retained by the purified joint representation shows that semantically matched visual information also contributes substantially to

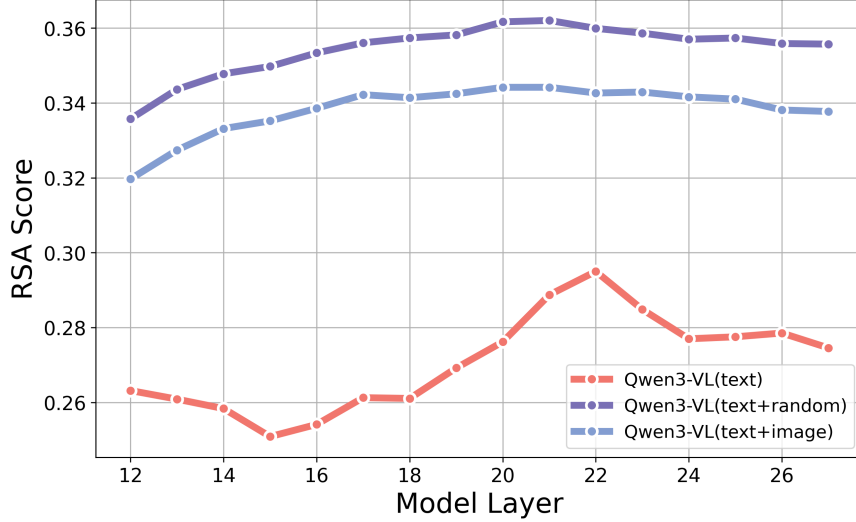


Figure 9: Representational Similarity Analysis (RSA) scores across middle-to-late model layers (Layers 12–27) of Qwen3-VL. The RSA score reflects the structural alignment between the model’s representational geometry and the brain’s neural response geometry. Both vision-augmented conditions substantially outperform the language-only baseline.

Table 5: Fine-grained RSA control analysis at layer 20.

Condition	RSA Score
Text + Noise	0.254
Text Only	0.278
Text + Matched Image	0.343
Text + Random Image	0.360
Purified Joint Representation	0.345

Table 6: Decodability of the targeted linear components before and after linear elimination. Values report five-fold cross-validated R^2 .

Target	Before	After
Text + Random Image	0.18	≈ 0
Text + Matched Image	0.44	≈ 0

the observed brain–representation alignment.

G Validation of Linear Elimination

The linear elimination procedure in Section 6.2 is designed to remove linearly decodable components associated with non-semantic visual effects while preserving the information specific to semantically matched visual inputs. To verify that the procedure operates as intended, we conducted two additional validation analyses.

Decodability of the Removed Components. We first examined whether the targeted linear components remain decodable after elimination. Ridge regression models were trained under five-fold cross-validation to predict the corresponding joint representations before and after the elimination procedure.

Before elimination, the relevant predictor explains approximately 18% of the variance in the *Text + Random Image* representation. After elimination, the cross-validated R^2 decreases to approximately zero. Similarly, for the matched-image

representation, the corresponding intermediate representation explains approximately 44% of the variance before elimination, while the cross-validated R^2 is reduced to approximately zero after elimination.

These results confirm that the linearly decodable components targeted by the elimination procedure are effectively removed.

Preservation of Semantic Structure. We further examined whether the purification procedure preserves the semantic structure of the representation rather than simply degrading the representation as a whole. We compared the representational geometry of the original and purified joint representations with that of the text-only representation.

The RSA score between *Text + Matched Image* and *Text Only* is 0.140, whereas the RSA score between the purified joint representation and *Text Only* increases to 0.256. This indicates that after removing the targeted non-semantic linear component, the remaining representation retains a stronger correspondence with the semantic geome-

try of the original textual representation.

Taken together, these validation analyses show that the linear elimination procedure effectively removes the targeted linearly decodable component while preserving, and relatively emphasizing, the semantic structure associated with the matched visual input.

H Experiments on LLaVA

We also conducted experiments with **LLaVA**. Its architecture is similar to **Qwen3-VL**, with **Qwen2** as the backbone language model. Across different experimental settings, LLaVA exhibits trends largely consistent with those observed for Qwen3-VL.

Our three main findings also hold for LLaVA. Compared with the Qwen3-VL series, we observe two notable differences (Figure 10): (1) When only text is provided as input, Qwen2 and LLaVA achieve lower brain scores in some regions, such as the Fusiform Face Area (FFA) and the Lateral Occipital Complex (LOC). (2) In the Superior Parietal Lobule (SPL), adding visual stimuli does not lead to a score improvement. This may be because the SPL is primarily involved in spatial perception and sensorimotor control, rather than visual-semantic processing.

I Experiments on LLaMA

We also conducted experiments with LLaMA, whose architecture differs substantially from the other two models. In Qwen-VL and LLaVA, the image representations encoded by the image encoder are concatenated with the text representations and fed into the first few layers of the backbone network for self-attention-based reasoning. In contrast, LLaMA injects the image representations in a layer-wise manner. Specifically, it evenly selects 8 positions in the LLM and inserts 8 network layers that perform cross-attention. Within these layers, text and image representations interact via cross-attention, while other layers continue with standard self-attention.

This design difference leads to the following experimental observations (Figure 11): brain scores fluctuate substantially, particularly when images with semantically matching content are input. Nevertheless, several layers still conform to our core findings, indicating that the LLaMA-Vision architecture is capable of producing joint representations that better predict neural responses to language. The structural difference also manifests in voxel-

level brain scores: when no images are provided, the improvement of multimodal models over unimodal models is limited in some regions. After adding images, certain regions, such as V5, show substantial increases, consistent with expectations.

J Brain Region Abbreviations and Descriptions

Table 7: Brain Region Abbreviations and Descriptions

Abbreviation	Description
STG	Superior Temporal Gyrus
TP	Temporal Pole
IFG (Op)	Inferior Frontal Gyrus (Opercular Part)
IFG (Tri)	Inferior Frontal Gyrus (Triangular Part)
V1	Primary Visual Cortex
V2	Secondary Visual Cortex
V3	Visual Area V3
V4	Visual Area V4
V5	Visual Area V5 (Middle Temporal Visual Area)
IT	Inferior Temporal Gyrus
FFA	Fusiform Face Area
PPA	Parahippocampal Place Area
LOC	Lateral Occipital Complex
IPS	Intraparietal Sulcus
SPL	Superior Parietal Lobule
AG	Angular Gyrus
SG	Supramarginal Gyrus

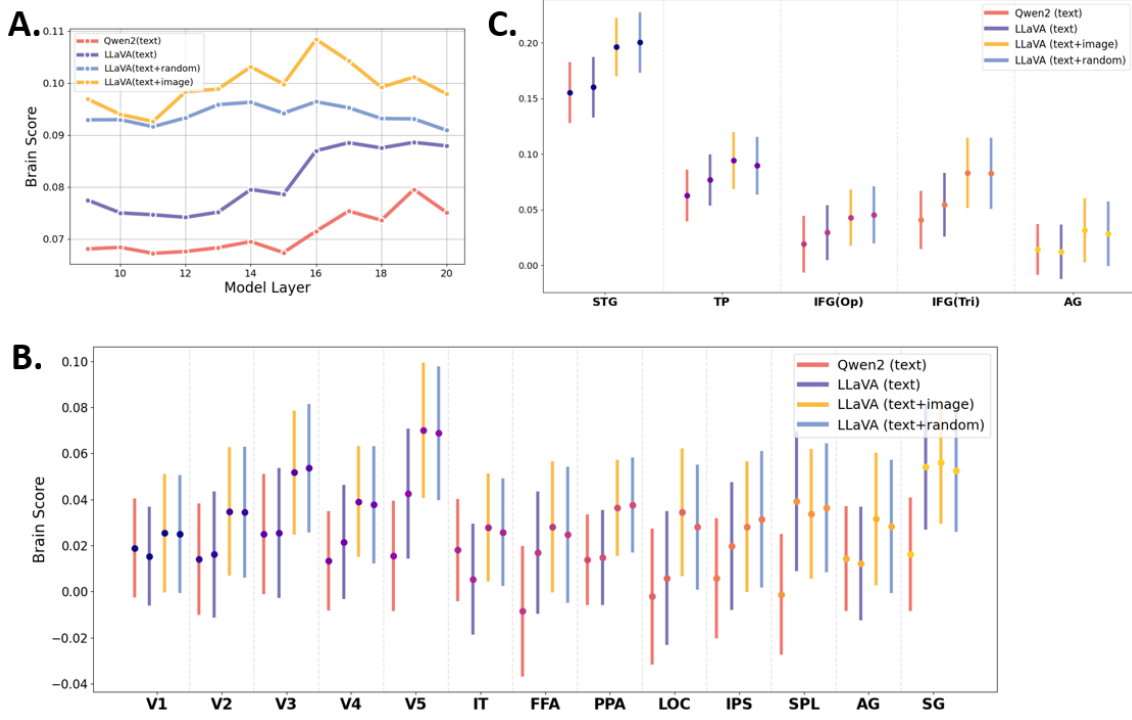


Figure 10: Experiments on LLaVA.

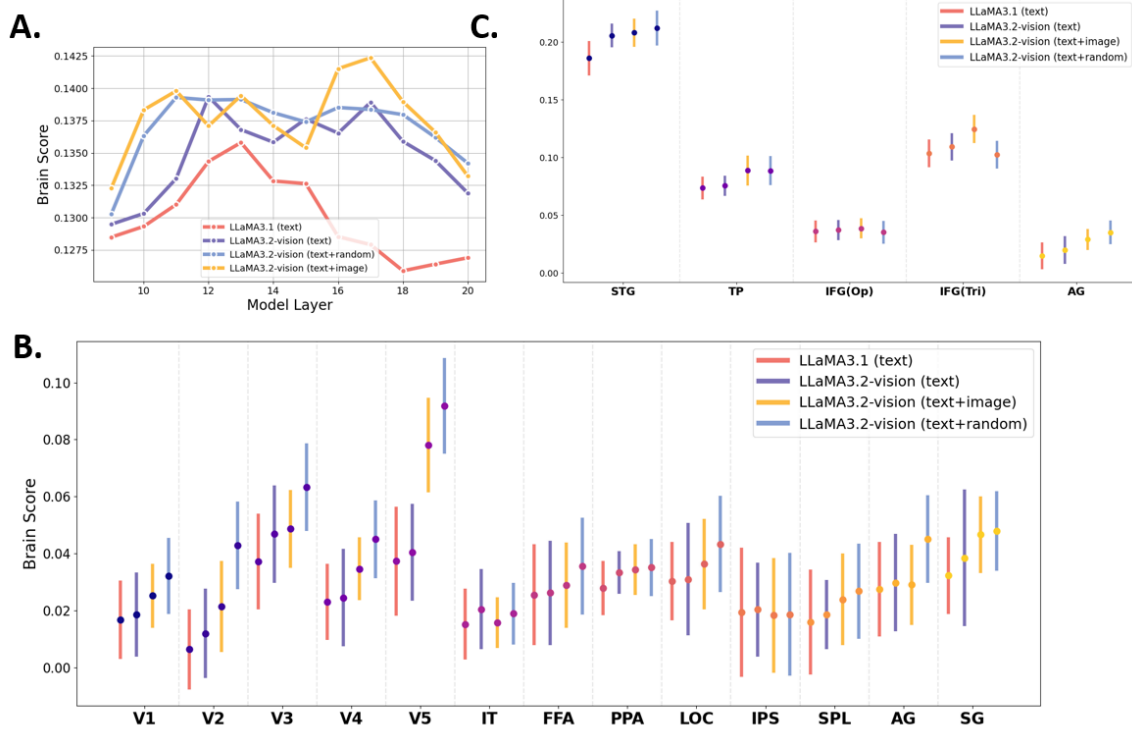


Figure 11: Experiments on LLaMA.