Vision-Language Models Align with Human Neural Representations in Concept Processing

Anna Bavaresco, Marianne de Heer Kloots, Sandro Pezzelle, Raquel Fernández
Institute for Logic, Language and Computation
University of Amsterdam
{a.bavaresco, m.l.s.deheerkloots, s.pezzelle, raquel.fernandez}@uva.nl

Abstract

Recent studies suggest that transformer-based vision-language models (VLMs) capture the multimodality of concept processing in the human brain. However, a systematic evaluation exploring different types of VLM architectures and the role played by visual and textual context is still lacking. Here, we analyse multiple VLMs employing different strategies to integrate visual and textual modalities, along with language-only counterparts. We measure the alignment between concept representations by models and existing (fMRI) brain responses to concept words presented in two experimental conditions, where either visual (pictures) or textual (sentences) context is provided. Our results reveal that VLMs outperform the language-only counterparts in both experimental conditions. However, controlled ablation studies show that only for some VLMs, such as LXMERT and IDEFICS2, brain alignment stems from genuinely learning more human-like concepts during *pretraining*, while others are highly sensitive to the context provided at *inference*. Additionally, we find that vision-language encoders are more brain-aligned than more recent, generative VLMs. Altogether, our study shows that VLMs align with human neural representations in concept processing, while highlighting differences among architectures. We open-source code and materials to reproduce our experiments at: https://github.com/dmg-illc/vl-concept-processing.

1 Introduction

Recent attempts to augment language-only models with vision have resulted in a multitude of vision-language models (VLMs), integrating modalities with different strategies based on the task at hand. While the practical implications of multimodality are evident—it allows language-only models to 'see' and perform novel vision-language tasks—, its theoretical implications from the point of view of human language modelling are not yet fully understood.

Insights from cognitive and neuroscientific work suggest that human semantic representations are deeply grounded in multimodal sensory experiences (Louwerse, 2011; Barsalou, 1999; Harnad, 1990; Bergen, 2012), and that visual and linguistic stimuli evoke shared neural activity patterns (Devereux et al., 2013; Simanova et al., 2014; Popham et al., 2021; Kaup et al., 2024). These findings motivate the hypothesis that semantic representations learnt by VLMs approximate the human ones better than representations by unimodal (both language-only and vision-only) models.

However, existing works testing this empirically point to diverging conclusions, with some studies documenting advantages of multimodality (e.g., Oota et al., 2022; Zhuang et al., 2024a), others suggesting that multimodal fine-tuning 'harms' the human-alignment properties originally acquired by language models (e.g., Bavaresco and Fernández, 2025), and more nuanced contributions indicating that the advantages of multimodality are circumscribed (e.g., Pezzelle et al., 2021; Dong and Toneva, 2023; Zhuang et al., 2024b; Ryskina et al., 2025). A likely reason behind the difficulties in reconciling the findings from these studies lies in their addressing similar research questions with different methods.

Among others, two under-investigated experimental factors that may strongly impact measures of human-likeness computed on VLM semantic representations concern the type and amount of context provided, and the differences between VLM architectures. More concretely, many psycholinguistic studies on multimodality in concept processing analysed human measures (both behavioural and neural) collected by presenting participants with non-contextualised words (e.g., Pezzelle et al., 2021; Zhuang et al., 2024b; Bavaresco and Fernández, 2025). This may have resulted in limited

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 3255–3274
March 24-29, 2026 ©2026 Association for Computational Linguistics

engagement of multimodal knowledge, ultimately leading to an under-appreciation of VLMs' brain modelling potential. An additional limitation of previous work is that it mostly focused on the multimodal/unimodal dichotomy, placing little emphasis on the substantial differences existing between multimodal architectures, and their implications for brain modelling.

In this study, we systematically compare ten models, including VLMs and language-only counterparts, by measuring their correlation with a subset of the Pereira dataset (Pereira et al., 2018), collecting human fMRI (functional magnetic resonance imaging) responses to 180 concepts. Our experiments address the limitations of previous work in two respects.

First, we explore the role played by the **context** provided through different modalities; we analyse brain responses to concept words that were presented to participants with either a sentential context (*sentence condition*, where concept words are presented within sentences) or a visual one (*picture condition*, where concept words are accompanied by images illustrating their content), as shown in Figure 1. Besides ensuring that multimodal knowledge is engaged, focusing on this dual-context setup allows us to better characterise the brain-aligning properties of VLMs: if VLMs are more brain-aligned than their language-only counterparts only when visual context is provided, their superiority may be simply due to accessing visual information that language-only models cannot 'see'; on the other hand, if VLMs have an advantage even without receiving an image input, this signals they have truly learnt human-like, multimodal concept representations during pretraining.

Second, we evaluate different **VLM families**: we indentify three classes of VLMs integrating the visual and textual modalities through different strategies, and test two representative exemplars for each. By comparing their alignment with brain responses, we test whether multimodality leads to a generalised pattern of improvement or is family/architecture-dependent.

Our results reveal that VLMs tend to be more brain-aligned than their language-only counterparts in both context conditions, suggesting that they successfully model human-like concept processing. In addition, we find that vision-language encoders, such as VisualBERT (Li et al., 2019) and LXMERT (Tan and Bansal, 2019), model brain responses better than the other VLM types, including the more powerful, generative ones. This means that performance on downstream tasks, where generative VLMs excel, may not go hand in hand with human-like concept processing. More broadly, our findings contribute to the ongoing debate around whether multimodality results in more human-like language models or not, showing that it is beneficial for brain alignment in concept processing.

2 Related Work

2.1 Multimodal models in cognitive modelling

Several recent studies have used transformer-based multimodal models in brain encoding/decoding experiments or to model behavioural data. Here, we review the closest to our focus, i.e., those analysing responses to concepts as opposed to more complex stimuli (e.g., videos).

Pezzelle et al. (2021) evaluated word representations extracted from several VLM architectures against human semantic judgments, including word similarity and relatedness benchmarks. Their findings reveal that VLM word representations align better with human judgments than representations from text-only models, although only on concrete word pairs. Similarly, Zhuang et al. (2024a) and Zhuang et al. (2024b) found visual grounding to result in improved and more efficient word learning, especially in low-data regimes.

Considering studies analysing brain activations, Oota et al. (2022) used several unimodal (language-only and vision-only) models and VLMs to predict a subset of the Pereira dataset (only the picture condition), and found a VLM (VisualBERT, Li et al., 2019) to be more brain-predictive than the other architectures. Crucially, in this setup, multimodal representations are the only ones receiving the same input shown to the human participants (i.e., image+word). Therefore, it is difficult to determine if their advantage is due to capturing something fundamental about concept processing, or simply to having access to more information than unimodal models.

Bavaresco and Fernández (2025) partially addressed this limitation by comparing multimodal and language-only models in a controlled setup, where they were all fed with text-only inputs and used to model fMRI responses collected while participants viewed isolated nouns. Their results, surprisingly, reveal an advantage of language-only models, even on a subset of more concrete nouns. As it is known that the presence of linguistic con-

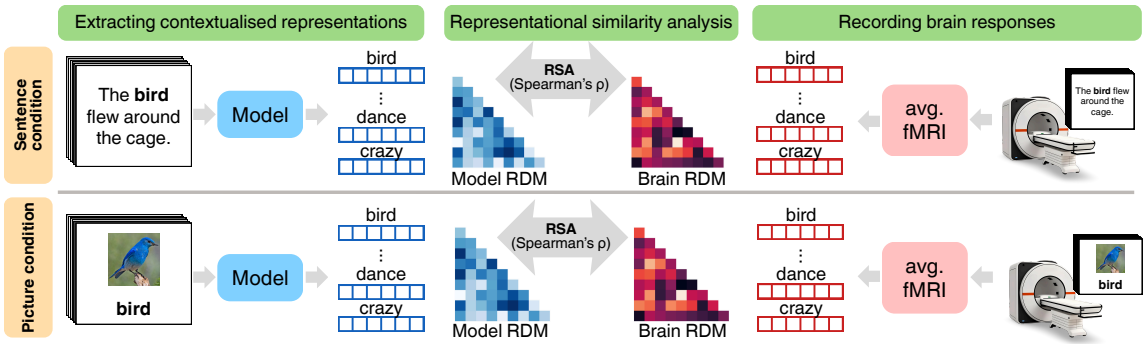

Figure 1: Overview of the experimental setup in the sentence (top) and picture (bottom) condition. Models are fed with the same stimuli participants saw in the fMRI scanner, i.e., concept words appearing in six contexts (provided by either sentences or pictures). Note that contexts are intended to highlight the same word meaning, but may describe different situations (sentences are *not* image captions). Model representations and brain responses averaged across the six contexts are then used to derive representational dissimilarity matrices (RDMs), storing pairwise cosine distances. Finally, the Spearman correlation between these RDMs provides a measure for model–brain alignment. Best viewed in colour.

text influences both brain responses to concepts (Xu et al., 2005; Deniz et al., 2023) and mental simulations of their content (Zwaan, 2014), a likely explanation for their findings is that the brain responses they analysed did not reflect a detectable engagement of multimodal knowledge.

Finally, Ryskina et al. (2025) analysed, again, the Pereira dataset and predicted fMRI responses using both VLMs and large language-only models (LLMs). Their main goal was to identify concept-sensitive brain regions and check whether their neural activations can be successfully modelled with VLMs and LLMs. Systematic comparisons of VLM families and analyses of the role played by context were, therefore, beyond the scope of their study.

We complement these works by analysing brain responses to concepts presented within different context conditions, including a picture condition, as analysed by Oota et al. (2022), and a sentence condition. This allows us to determine if VLMs' alignment with human brain responses is modulated by the context or consistent across visual and sentential contexts.

2.2 VLM families

Transformer-based vision-language models (VLMs) can be divided into three categories: contrastive models, vision-language encoders, and generative VLMs. Contrastive models, such as CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), and LiT (Zhai et al., 2022), encode images and text separately, with two dedicated transformer-based modules. These modules are pre-trained with a contrastive loss, which maximises the similarity between image and text embeddings of matching image-text pairs.

Vision-language encoders are characterised by a specific module that uses attention mechanisms to learn relations between visual features extracted with an object detector and text embeddings. In some architectures, such as VisualBERT (Li et al., 2019), language processing and multimodal integration are performed by the same BERT-based (Devlin et al., 2019) module; in other architectures, e.g., LXMERT (Tan and Bansal, 2019), text is encoded in a dedicated transformer module before being passed to the cross-modal module.

Lastly, generative VLMs consist of a pretrained LLM, a pretrained vision encoder, and an 'adaptor' (or 'projector'), i.e., a shallow module which learns a mapping between the space of the image tokens and that of the language tokens. Examples of these architectures are LLaVA-NeXT (Liu et al., 2024), IDEFICS2 (Laurençon et al., 2024), Qwen2.5-VL (Bai et al., 2025), and Molmo (Deitke et al., 2024).[1]

The studies reviewed above (Section 2.1) used exemplars from at most two of the VLM families, but a comprehensive comparison remains elusive. To address this gap, we experiment with six different VLMs, including representatives from each family.

[1]Here, we do not review proprietary models as they do not allow extracting layer-wise representations, which are fundamental for assessing brain alignment.

3 Methods

To study the alignment between vision-language models (VLMs) and brain responses, we focus on a dataset of neural responses to concept words, which participants read either accompanied by pictures or within sentences (Experiment 1 in Pereira et al., 2018). We derive representations from the same concept words using a set of VLMs and language-only models and quantify their alignment with human responses in two brain networks with representational similarity analysis (RSA, Kriegeskorte et al., 2008). An overview of our experimental setup is provided in Figure 1.

3.1 Brain responses

The brain responses we focus on were collected by Pereira et al. (2018). They consist of voxel-wise fMRI activations collected while 16 participants were presented with English words representing specific concepts in different conditions. There are 180 concept words in total, including different parts of speech (128 nouns, 22 verbs, 29 adjectives and adverbs, and 1 function word; see Appendix A.1 for the full list).

We consider two experimental conditions: a language-only *sentence condition* and a multimodal *picture condition*. In the sentence condition, participants were fMRI-scanned while reading sentences where the target words were boldfaced. For each concept word, participants saw six sentences, one at a time. In the picture condition, each word was presented together with an image illustrating the relevant concept. Again, participants viewed six different images for each concept word, one at a time. In both conditions, participants were asked to think about the meaning of the target concept.

We use the brain responses as preprocessed by Pereira et al. (2018). More concretely, the preprocessed response for each stimulus consists of an array where entries represent the 'magnitude' of the brain activity at different voxels. While responses were recorded for the whole brain, we focus on two specific regions, involved in either linguistic or visual processing: the left-hemisphere *Language network* (Fedorenko et al., 2011) and the *Visual network* (Power et al., 2011; Buckner et al., 2008). Responses are averaged across the six presentations of the same word per condition; hence, for each participant, we have one averaged brain response per concept, condition, and brain network. See Appendix A.2 for additional details on the brain responses and the anatomical regions included in the two brain networks.

3.2 Models

We employ two main types of models: a set of VLMs trained on images and text, and a set of language-only models trained exclusively on text. To make our evaluation comprehensive, we include multiple VLMs that are representative of the families described in Section 2. Regarding the language-only models, we include architectures that either provide informative baselines (e.g., GloVe) or useful comparisons with specific VLMs.[2] We briefly describe the models here and refer to Appendix B.1 for details about the specific implementations.

Vision-language models We select two representative VLMs from each family. More specifically, we consider CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) for the contrastive VLMs, VisualBERT (Li et al., 2019) and LXMERT (Tan and Bansal, 2019) for the vision-language encoders, and IDEFICS2 (Laurençon et al., 2024) and LLaVA NeXT (Liu et al., 2024) for the generative VLMs.

Language-only models We experiment with one encoder-only language model (BERT, Devlin et al., 2019) and two decoder-only large language models—Mistral (Jiang et al., 2023) and Llama3 (Grattafiori et al., 2024). In addition, we include the simpler distributional semantic model GloVe (Pennington et al., 2014). BERT forms the basis of the language components of ALIGN, VisualBERT and LXMERT,[3] while Mistral and Llama3 are the language models used in IDEFICS2 and LLaVA NeXT, respectively. Finally, while GloVe does not output contextualised representations, its embeddings were used by Pereira et al. (2018) to select the concept stimuli for the fMRI study; we therefore include them as they provide a useful reference.

3.3 Extracting representations

To extract model representations that we can compare to the fMRI responses, we feed the models with the same stimuli presented to the participants in the experiments by Pereira et al. (2018). Different models require slightly different inputs, as we

[2]Note that testing state-of-the-art models is not crucial for our goals. We choose fully open-source VLMs that have a language-only counterpart and were among the best ones when the experiments were conducted.

[3]While VisualBERT is initialised with BERT weights, LXMERT and ALIGN train their BERT-based modules from scratch.

explain in detail in this section.

Sentence condition In this condition, participants read sentences without visual information. All transformer-based models (both VLMs and language-only models) are fed with the same sentences, without any image. One exception to the procedure is LXMERT; as it requires some visual input in addition to text, we pass an image made up of random noise along with each sentence.[4]

After extracting model representations (*hidden states*) for the entire sentences from each model layer, we select those corresponding to the tokens of the target concept word. If the target word consists of several tokens, we average the hidden states of the relevant tokens. Finally, we average again over the six contexts where the word appears. All averages are computed layer-wise. This means that our procedure yields one concept representation for each concept at each model layer.

As for GLoVe, given that it provides static (type-based) decontextualised representations, we simply select the pre-trained GloVe embedding corresponding to each target word.

Picture condition In this condition, participants see each word together with an image. For the VLMs, both the target word and each image accompanying it are fed to the models. When using contrastive models, we integrate the image and text embeddings by taking their element-wise sum.[5] For the VL encoders, we extract hidden states from all text-specific layers and from the cross-modal layers. In generative VLMs, we consider the hidden states from all layers of the LLM module. To obtain a final representation for each concept word that can be compared to brain activations, we then average the extracted representations across the six images per word. For the language-only models, only the target word is provided as input.

3.4 Evaluation

We evaluate the models' alignment to fMRI data with Representational Similarity Analysis (RSA, Kriegeskorte et al., 2008), which is illustrated in Figure 1. In the following, we explain how it differs from other existing metrics and describe how it was computed in our setup.

[4]We create a different random-noise image for each sentence. These images are linked to our public GitHub repository.

[5]We also tested element-wise multiplication and concatenation, but these resulted in less brain-aligned representations.

RSA vs. other metrics Given two representational spaces (e.g., those defined by fMRI responses and model representations), RSA measures whether the between-stimuli relations (typically distances) in one space are similar to the relations in the other (*second-order isomorphism*, Shepard and Chipman, 1970). This means that the information it provides is at the level of similarity between representational geometries (*representational similarity*), rather than between feature spaces (Dujmovic et al., 2024).

Alternative metrics that are better suited to study brain predictivity at the feature level concern the 'brain encoding' performance, or 'linear predictivity' (Naselaris et al., 2011). These methods involve training transformations from the model embedding space to optimise the fit with neural data, which usually results in higher correlation scores than RSA. However, such metrics have been found lacking in other desirable properties such as functional consistency (e.g., Bo et al., 2025). Additionally, the fact that they involve an optimisation step makes it difficult to disentangle how much the resulting correlations are attributable to model features vs. optimisation.

Given that the best choice of alignment metric ultimately depends on the research goals (Ivanova et al., 2021a), we deem RSA to be suitable for our interest in observing relative differences between models rather than maximising neural predictivity.

Computing RSA As routinely done in RSA, we approximate brain and model spaces through representational dissimilarity matrices (RDMs), whose entries store cosine distances between concept representations (i.e., brain responses or model embeddings). We then quantify their alignment as a Spearman correlation over the vectorised off-diagonal elements of the RDMs, corresponding to the unique set of pairwise distances in each space. A significant positive correlation between RDMs indicates that there is representational similarity between brain-derived and model-derived spaces. We compute representational similarity against brain RDMs averaged across single participants.[6] For each model with multiple layers, RSA is computed separately for each layer's representations. When drawing comparisons between models, we consider the best (i.e., most aligned) layer for each model. We verify that differences in brain alignment be-

[6]We elaborate more on the reasons behind this choice in the Limitations section at the end of the paper.

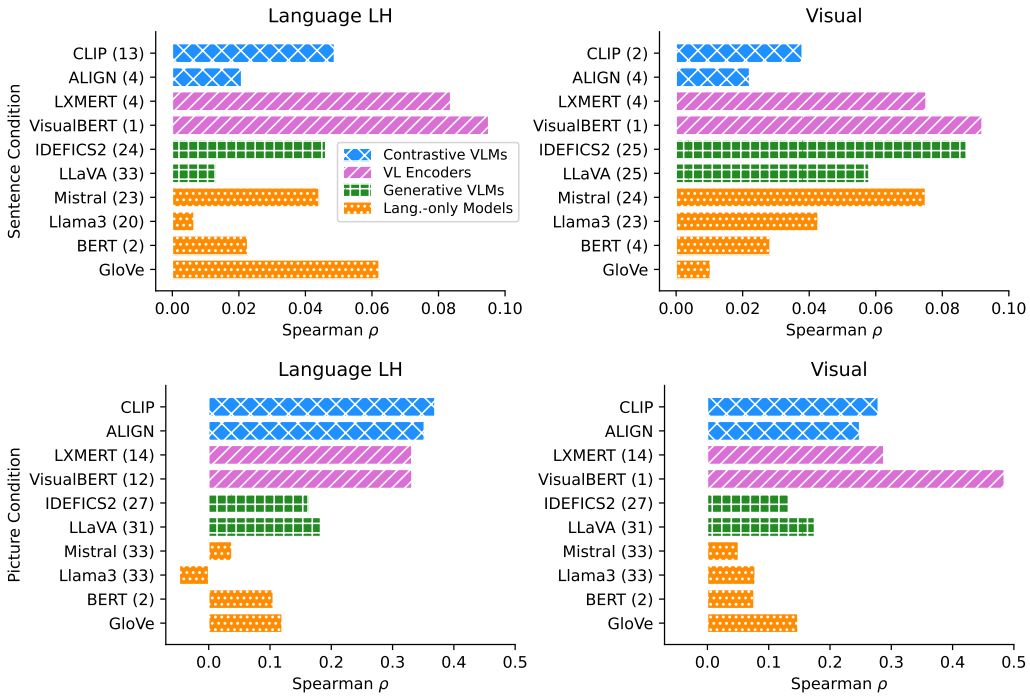

Figure 2: RSA results for the *sentence condition* (upper row) and *picture condition* (lower row). Spearman correlations indicate the alignment between concept representations by models and fMRI responses in the left-hemisphere (LH) language network and in the visual network. Numbers in brackets indicate the model layer from which representations were extracted. Note that the range of the x axes differs between conditions.

tween models are statistically significant by applying a Fisher transformation to all the unique pairs of Spearman correlations and calculating the p-value associated with the difference between the two z-scores. To control for false positives due to multiple comparisons (45 per brain network in each condition), p-values are Bonferroni-corrected with $\alpha = 0.05$. We report the complete set of p-values in Appendix C.1, Table 2. To obtain a random baseline for the representational similarity, we compute alignment against shuffled brain responses, i.e., a condition where each concept was associated with a randomly chosen fMRI response from a non-matching concept.

4 Results

Our main results are visualised in Figure 2, which reports representational similarity from the most brain-aligned model layer in the cases where we extract representations from multiple. All the displayed correlations are significantly different from 0, which coincides with the random baseline computed by shuffling brain responses. Layer-wise RSA values for all models are provided in Appendix C.2.

Sentence condition As displayed in the upper row of Figure 2, all models exhibit low to moderate positive correlations with brain responses from both networks. In the language network, the VL encoder VisualBERT is statistically significantly more brain-aligned than all the other models ($\rho = 0.10$) except for LXMERT ($\rho = 0.08$). Additionally, GloVe's correlation with fMRI responses is not significantly different from that of the more advanced models—both multimodal and unimodal—CLIP, LXMERT, IDEFICS2 and Mistral. Comparisons between VLMs and their language-only counterparts reveal that both VisualBERT and LXMERT are significantly more brain-aligned than BERT, while LLaVA and IDEFICS2 are not different from their language decoders Llama3 and Mistral.

In the visual network, language-only models surprisingly achieve significant correlations, with Mistral performing comparably with the VLM LXMERT and outperforming LLaVA, CLIP and ALIGN. However, the highest correlations are still observed for the VLMs VisualBERT and IDEFICS2 ($\rho = 0.09$ for both), which significantly outperform all other models except for LXMERT and Mistral. Regarding the remaining comparisons between VLMs and language-only counterparts,

LLaVA's advantage over Llama3 is not significant, while both VisualBERT and LXMERT are significantly more brain-aligned than BERT. Remarkably, language-only models achieve significant correlations in this network even if no visual information is being presented.

Picture condition Results displayed in the lower row of Figure 2 show higher correlations than in the sentence condition, which can be attributed to higher signal (i.e., higher inter-participant similarities) in the fMRI responses. Notably, VLMs are significantly more brain-aligned than their unimodal counterparts in both brain networks.

In the LH language network, both contrastive VLMs and VL encoders exhibit moderate correlations with brain responses ($\rho > 0.3$), which are all statistically significantly stronger than those achieved by generative VLMs and language-only models. In the visual network, VisualBERT outperforms all other models ($\rho = 0.48$) and CLIP, ALIGN and LXMERT exhibit similar brain correlations ($0.25 < \rho < 0.29$).

Lastly, GloVe exhibits higher correlations than all other more powerful language-only models in the language network. This is likely because Pereira et al. (2018) selected the concept stimuli to include in their experiment using GloVe representations, as we mentioned above.

4.1 Trends across the board

Overall, our results indicate an advantage of VLMs over the language-only counterparts that is consistent across conditions and brain networks. Below, we highlight the main implications regarding differences among VLMs and comparisons with their unimodal counterparts.

VLM families Our findings reveal that VL encoders are the most brain-aligned, outperforming the other model families in all scenarios except for the language network in the picture condition, where contrastive VLMs have an advantage. A possible explanation for this result is that contrastive VLMs are effective at capturing object-word correspondences—a scenario quite akin to the picture condition—but struggle to accurately represent more complex relations between entities (Thrush et al., 2022; Liu et al., 2023; Parcalabescu et al., 2022; Hendricks and Nematzadeh, 2021). This limitation may, therefore, have resulted in lower performance compared to the VL encoders.

Regarding generative VLMs, their brain correlations in the picture condition are *lower* than those by contrastive VLMs and VL encoders. Even in the sentence condition, where there is linguistic context they can incorporate in their word representations, they still do not outperform VL encoders—a surprising result, given their superior performance on downstream tasks. A potential reason for this may be that autoregressive pretraining privileges production over representation (BehnamGhader et al., 2024; Muennighoff et al., 2025; Springer et al., 2024), making *word*-level representations less expressive than those by previous architectures.

VLMs vs. language-only counterparts In general, we observe an advantage of VLMs over language-only models, albeit with differences between conditions. In the picture condition, all VLMs are more brain-aligned than their unimodal counterparts, indicating that their ability to process visual information is beneficial for modelling brain-relevant semantic aspects.

In the sentence condition, results are more nuanced, with VisualBERT and LXMERT outperforming BERT, and IDEFICS2, LLaVA and ALIGN never significantly outperforming Mistral, Llama3 and BERT, respectively. This suggests that the type of context provided (visual vs. sentential) affects the brain-modelling abilities of VLMs vs. their language-only counterparts. We further investigate this aspect in two ablation studies.

5 Ablation Studies

To complement the findings provided by RSA, we conduct two ablation analyses aimed at better understanding what drives VLMs' brain alignment in both experimental conditions.

5.1 Semantic information in the sentence condition

The analyses presented earlier allow comparing the brain alignment of VLMs against that of the language-only counterparts. However, finding that VLMs achieve similar brain alignment to unimodal counterparts does not automatically indicate they rely on the same knowledge: while it seems natural to assume that representations by a VLM will incorporate the brain-relevant information by the underlying language module (and perhaps learn additional information as a result of multimodal fine-tuning), it is still possible that vision-language fine-tuning alters the language-module information substantially. The two scenarios are illustrated in

Appendix B.2, Figure 5. To investigate this, we conduct a partial correlation analysis aimed at removing from VLMs' embedding spaces the information shared with LLMs' spaces. We focus on a subset of models whose architecture is highly similar to that of the language-only counterpart, i.e., VisualBERT, LXMERT, LLaVA and IDEFICS2.[7]

Computing partial correlations Partial correlations are computed as follows. Consider the RDM from a VLM y, the RDM from its language-only counterpart x, and the residuals $r_i = y_i - \hat{y}_i$ from the linear regression with equation $\hat{y}_i = a + bx_i$. The partial correlation achieved by the VLM is defined as the Spearman correlation $\rho(r_i, z_i)$, where z represents the RDM of the brain responses, and it provides an indication of the VLM's brain alignment achieved once the representational information shared with the language-only counterpart is ablated.

Results VLMs' initial (as reported in the main experiment) and partial correlations with brain responses are displayed in Figure 3. If the brain alignment of a VLM is driven by its language-only component, the partial correlation should be significantly lower than the initial correlation, signalling that the removed information mattered. In contrast, the absence of such a difference indicates that the semantic knowledge learnt by the language-only module was not responsible for the brain alignment achieved by the VLM; hence, the knowledge driving alignment must come from multimodal training.

In the language network, all differences between partial and initial correlations—except for IDEFICS2 in language LH—are *not* statistically significant, suggesting that the brain alignment achieved by these models cannot be attributed to semantic information already present in the language-only modules, but to different knowledge acquired during multimodal pretraining. In the visual network, results reveal diverging patterns between generative VLMs and the other VLMs: the differences between initial and partial correlations are statistically significant for the former, while they are not for the latter. Interestingly, this suggests that part of the information relevant for alignment with visual brain responses was already present in Mistral and

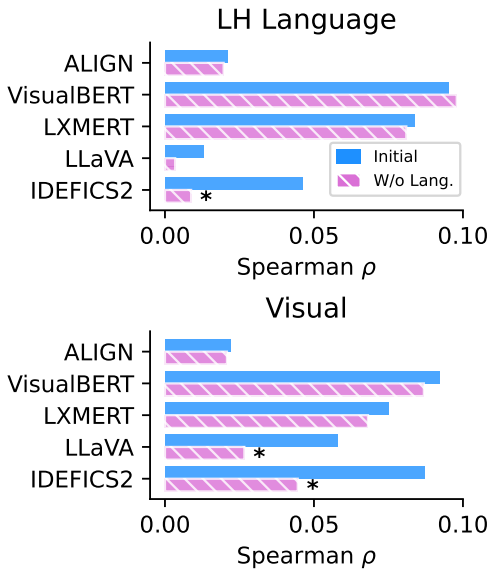

Figure 3: Initial (as reported in the main experiment) and partial correlations between VLM representations and fMRI responses in the *sentence condition*. Statistically significant differences (marked by asterisks) between initial and partial correlations indicate that the brain-relevant information captured by the VLM is shared with that present in its language module.

Llama3 before any vision-language training. That is, the brain-aligning information in these language-only models (see Figure 2) seems to be maintained after multimodal fine-tuning.

5.2 Visual information in the picture condition

In the picture condition, we found a systematic advantage of VLMs over their language-only counterparts. However, a potential confound could be that VLMs have access to additional contextual information—the picture—that is available to human participants but not to language-only models. In this sense, their superior alignment with human responses could stem from an uneven comparison rather than indicate that they capture additional semantic information.

To shed more light on this issue, we conduct an ablation study where we pass only the concept word (without pictures[8]) to the VLMs, so that the input they receive matches that provided to language-only models. We then recompute RSA against human fMRI responses from the picture condition, following the same procedures employed in the main experiment. The resulting Spear-

[7]While VisualBERT, LLaVA and IDEFICS2 were initialised with pretrained weights from their language-only counterparts, ALIGN and LXMERT were trained multimodally from scratch and are hence not included.

[8]Again, LXMERT requires a visual input. As we do in the sentence condition, we pass a random-noise image.

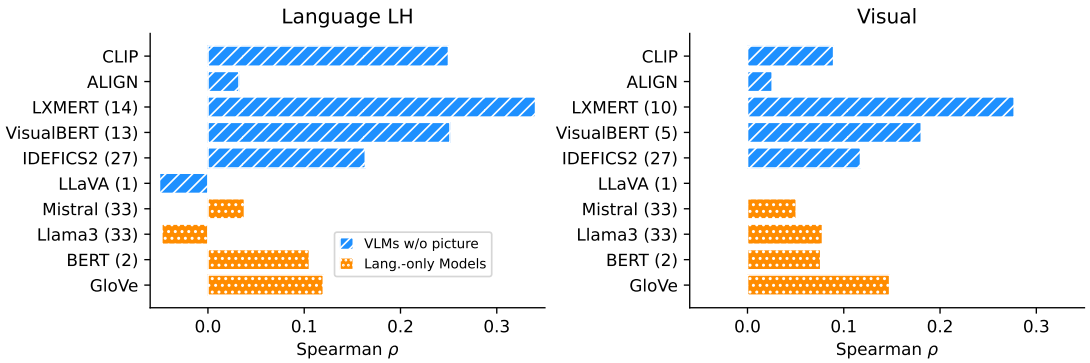

Figure 4: Results from the ablation study where we pass only concept words to both VLMs and language-only models. For both brain networks, we show the Spearman correlations resulting from RSA, indicating the alignment between models and fMRI responses from the *picture condition*. Numbers in brackets indicate the layers from which representations are extracted.

man correlations are provided in Figure 4. We assess the statistical significance of differences between each pair of correlations per network as in the main experiment.

This analysis reveals striking differences between VLMs: CLIP, ALIGN, VisualBERT, and LLaVA suffer dramatic drops in brain alignment when the picture is not provided as input, indicating their initial correlation was highly influenced by the context provided by the picture; on the other hand, brain-alignment variations observed for LXMERT and IDEFICS2 are minimal, suggesting their brain alignment is less sensitive to the presence of an input image (see also Figure 6 in the Appendix).

Despite the performance drop suffered by some of the VLMs, the most brain-aligned architectures in both networks remain multimodal: LXMERT is significantly more brain-aligned than all other models, and VisualBERT significantly outperforms all language-only models.

Considering VLMs and their language-only counterparts, LXMERT and VisualBERT significantly outperform BERT, while ALIGN is *less* brain-aligned than BERT across both brain networks. As for the generative VLMs, IDEFICS2 remains significantly more brain-aligned than Mistral, and LLaVA is significantly *less* brain-aligned than LLama3 in the visual network.

Overall, this analysis suggests that, while some architectures rely heavily on input images, others yield strong brain correlations even without meaningful visual input.

6 Conclusion

We provide a broad investigation of the brain alignment to human concept processing achieved by vision-language models from different families, drawing meaningful comparisons with language-only models and considering two experimental conditions and two brain networks.

Our results show evidence that the highest brain alignment is consistently achieved by one of the VLMs (and not a language-only model), although not the same architecture across all conditions and brain networks. Additionally, we find that vision-language encoders tend to exhibit higher brain alignment than the more recent generative VLMs. Lastly, our findings demonstrate that the superior brain alignment achieved by vision-language encoders cannot be solely attributed to receiving additional input at inference, but stems from learning novel multimodal semantic information during their pretraining.

More broadly, our findings speak to a wider research effort aimed at better understanding the effects of multimodality on language modelling. This includes works showing that vision-language training harms performance on natural-language-understanding (Iki and Aizawa, 2021) and commonsense-reasoning benchmarks (Madasu and Lal, 2023), and others finding it results in improved deployment of taxonomic knowledge (Qin et al., 2025) and prediction of words in context (Wang et al., 2023), but without fundamentally altering linguistic representations. We believe that our work, together with concurrent NeuroAI studies such as those reviewed in the previous sections, contributes an additional meaningful perspective, and we hope it will inspire future efforts to evaluate and design more human-like VLMs.

Limitations

Our aim was to provide a comprehensive evaluation of VLMs off-the-shelf, as they are being used by the AI community. It is important to consider that these VLMs differ along many dimensions, including size, architecture, learning objective and amount of training data. While the differences in brain alignment we observed between them are interesting and meaningful, our setup does not allow attributing them to one specific factor.

Another potential limitation of our study pertains to the ROIs we considered. The visual network and the left-hemisphere language network we analysed provide useful insights into how the brain responds to visuo-linguistic stimuli, but two caveats remain. The first is that the involvement of the language network in semantic processing is still under investigation, with recent studies suggesting that its activation may not be required for semantic processing (Ivanova et al., 2021b), and that additional regions outside this network are activated by concepts irrespective of the modality through which they are presented (Ryskina et al., 2025). Relatedly, the second caveat is that part of the multimodal integration arising from concept processing may happen in brain regions that we did not analyse, but were previously shown to be involved in multimodal processing (Fairhall and Caramazza, 2013; Bonner et al., 2013; Handjaras et al., 2016; Nikolaus et al., 2024).

Finally, we did not investigate inter-participant differences in brain alignment. We chose to average brain responses (more precisely, *brain RDMs*) across participants due to the high amount of noise in the individual data, especially in the language network. Averaging responses allowed us to get a stronger signal when measuring correlations with model representations. At the same time, this procedure may have 'evened out' some participant-specific patterns.

Acknowledgments

We thank the members of the Dialogue Modelling Group (DMG) from the University of Amsterdam and Leonardo Bertolazzi for the helpful feedback provided at different stages of this project.

We express our gratitude to the anonymous reviewers who offered valuable insights on the current and previous submissions. Their advice has contributed to improving the present work significantly.

A. B. and R. F. are funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 819455). M. d. H. K. is funded by the Netherlands Organisation for Scientific Research (NWO), through a Gravitation Grant 024.001.006 to the Language in Interaction Consortium.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*.

Lawrence W Barsalou. 1999. Perceptual symbol systems. *Behavioral and brain sciences*, 22(4):577–660.

Anna Bavaresco and Raquel Fernández. 2025. Experiential semantic information and brain alignment: Are multimodal models better than language models? In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 141–155, Vienna, Austria. Association for Computational Linguistics.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. LLM2Vec: Large Language Models Are Secretly Powerful Text Encoders. In *First Conference on Language Modeling*.

Benjamin K Bergen. 2012. *Louder than words: The new science of how the mind makes meaning*. Basic Books.

Yiqing Bo, Ansh Soni, Sudhanshu Srivastava, and Meenakshi Khosla. 2025. Evaluating Representational Similarity Measures from the Lens of Functional Correspondence. Proceedings of the Cognitive Computational Neuroscience conference (CCN).

Michael F Bonner, Jonathan E Peelle, Philip A Cook, and Murray Grossman. 2013. Heteromodal conceptual processing in the angular gyrus. *Neuroimage*, 71:175–186.

Randy L Buckner, Jessica R Andrews-Hanna, and Daniel L Schacter. 2008. The brain's default network: anatomy, function, and relevance to disease. *Annals of the new York Academy of Sciences*, 1124(1):1–38.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, and 1 others. 2024. Molmo and PixMo: Open weights and open data for state-of-the-art multimodal models. *CoRR*.

Fatma Deniz, Christine Tseng, Leila Wehbe, Tom Dupré la Tour, and Jack L Gallant. 2023. Semantic representations during language comprehension are affected

by context. *Journal of Neuroscience*, 43(17):3144–3158.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dota Tianai Dong and Mariya Toneva. 2023. Vision-language integration in multimodal video transformers (partially) aligns with the brain. *arXiv preprint arXiv:2311.07766*.

Marin Dujmovic, Jeffrey Bowers, Federico Adolfi, and Gaurav Malhotra. 2024. Inferring DNN-brain alignment using representational similarity analyses can be problematic. In *ICLR 2024 Workshop on Representational Alignment*.

Scott L Fairhall and Alfonso Caramazza. 2013. Brain Regions that Represent Amodal Conceptual Knowledge. *Journal of Neuroscience*, 33(25):10552–10558.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Giacomo Handjaras, Emiliano Ricciardi, Andrea Leo, Alessandro Lenci, Luca Cecchetti, Mirco Cosottini, Giovanna Marotta, and Pietro Pietrini. 2016. How concepts are encoded in the human brain: a modality independent, category-based cortical organization of semantic knowledge. *Neuroimage*, 135:232–242.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.

Lisa Anne Hendricks and Aida Nematzadeh. 2021. Probing Image-Language Transformers for Verb Understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3635–3644, Online. Association for Computational Linguistics.

Taichi Iki and Akiko Aizawa. 2021. Effect of visual extensions on natural language understanding in vision-and-language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2189–2196, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Anna A. Ivanova, John Hewitt, and Noga Zaslavsky. 2021a. Probing artificial neural networks: insights from neuroscience. ICLR 2021 Workshop: How Can Findings About The Brain Improve AI Systems?

Anna A Ivanova, Zachary Mineroff, Vitor Zimmerer, Nancy Kanwisher, Rosemary Varley, and Evelina Fedorenko. 2021b. The language network is recruited but not required for nonverbal event semantics. *Neurobiology of Language*, 2(2):176–201.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Barbara Kaup, Rolf Ulrich, Karin M Bausenhart, Donna Bryce, Martin V Butz, David Dignath, Carolin Dudschig, Volker H Franz, Claudia Friedrich, Caterina Gawrilow, and 1 others. 2024. Modal and amodal cognition: An overarching principle in various domains of psychology. *Psychological Research*, 88(2):307–337.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. 2008. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*.

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*.

Fangyu Liu, Guy Emerson, and Nigel Collier. 2023. Visual Spatial Reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Max M. Louwerse. 2011. Symbol Interdependency in Symbolic and Embodied Cognition. *Topics in Cognitive Science*, 3(2):273–302. _eprint:

https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1756-8765.2010.01106.x.

Avinash Madasu and Vasudev Lal. 2023. Is multimodal vision supervision beneficial to language? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2637–2642.

Niklas Muennighoff, Hongjin SU, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2025. Generative representational instruction tuning. In *The Thirteenth International Conference on Learning Representations*.

Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. 2011. Encoding and decoding in fMRI. *Neuroimage*, 56(2):400–410.

Mitja Nikolaus, Milad Mozafari, Nicholas Asher, Leila Reddy, and Rufin VanRullen. 2024. Modality-agnostic fmri decoding of vision and language. In *ICLR 2024 workshop on Representational Alignment (Re-Align)*.

Subba Reddy Oota, Jashn Arora, Veeral Agarwal, Mounika Marreddy, Manish Gupta, and Bapi Surampudi. 2022. Neural language taskonomy: Which NLP tasks are the most predictive of fMRI brain activity? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3220–3237, Seattle, United States. Association for Computational Linguistics.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2022. VALSE: A Task-Independent Benchmark for Vision and Language Models Centered on Linguistic Phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8253–8280, Dublin, Ireland. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J. Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. 2018. Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, 9(1):963.

Sandro Pezzelle, Ece Takmaz, and Raquel Fernández. 2021. Word Representation Learning in Multimodal Pre-Trained Transformers: An Intrinsic Evaluation. *Transactions of the Association for Computational Linguistics*, 9:1563–1579.

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.

Jonathan D Power, Alexander L Cohen, Steven M Nelson, Gagan S Wig, Kelly Anne Barnes, Jessica A Church, Alecia C Vogel, Timothy O Laumann, Fran M Miezin, Bradley L Schlaggar, and 1 others. 2011. Functional network organization of the human brain. *Neuron*, 72(4):665–678.

Yulu Qin, Dheeraj Varghese, Adam Dahlgren Lindström, Lucia Donatelli, Kanishka Misra, and Najoung Kim. 2025. Vision-and-language training helps deploy taxonomic knowledge but does not fundamentally alter it. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Maria Ryskina, Greta Tuckute, Alexander Fung, Ashley Malkin, and Evelina Fedorenko. 2025. Language models align with brain regions that represent concepts across modalities. In *Second Conference on Language Modeling*.

Roger N Shepard and Susan Chipman. 1970. Second-order isomorphism of internal representations: Shapes of states. *Cognitive psychology*, 1(1):1–17.

Irina Simanova, Peter Hagoort, Robert Oostenveld, and Marcel A. J. van Gerven. 2014. Modality-Independent Decoding of Semantic Information from the Human Brain. *Cerebral Cortex*, 24(2):426–434.

Jacob M. Springer, Suhas Kotha, Daniel Fried, Graham Neubig, and Aditi Raghunathan. 2024. Repetition improves language model embeddings. *ArXiv*, abs/2402.15449.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. 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 5100–5111.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5238–5248.

Wentao Wang, Wai Keen Vong, Najoung Kim, and Brenden M Lake. 2023. Finding structure in one child's linguistic experience. *Cognitive science*, 47(6):e13305.

Jiang Xu, Stefan Kemeny, Grace Park, Carol Frattali, and Allen Braun. 2005. Language in context: Emergent features of word, sentence, and narrative comprehension. *Neuroimage*, 25(3):1002–1015.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. 2022. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18123–18133.

Chengxu Zhuang, Evelina Fedorenko, and Jacob Andreas. 2024a. Lexicon-level contrastive visual-grounding improves language modeling. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 231–247, Bangkok, Thailand. Association for Computational Linguistics.

Chengxu Zhuang, Evelina Fedorenko, and Jacob Andreas. 2024b. Visual grounding helps learn word meanings in low-data regimes. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1311–1329, Mexico City, Mexico. Association for Computational Linguistics.

Rolf A Zwaan. 2014. Embodiment and language comprehension: Reframing the discussion. *Trends in cognitive sciences*, 18(5):229–234.

A Data

A.1 Concepts

The full list of concept words from the Pereira dataset (Pereira et al., 2018) is the following:

> Ability, Accomplished, Angry, Apartment, Applause, Argument, Argumentatively, Art, Attitude, Bag, Ball, Bar, Bear, Beat, Bed, Beer, Big, Bird, Blood, Body, Brain, Broken, Building, Burn, Business, Camera, Carefully, Challenge, Charity, Charming, Clothes, Cockroach, Code, Collection, Computer, Construction, Cook, Counting, Crazy, Damage, Dance, Dangerous, Deceive, Dedication, Deliberately, Delivery, Dessert, Device, Dig, Dinner, Disease, Dissolve, Disturb, Do, Doctor, Dog, Dressing, Driver, Economy, Election, Electron, Elegance, Emotion, Emotionally, Engine, Event, Experiment, Extremely, Feeling, Fight, Fish, Flow, Food, Garbage, Gold, Great, Gun, Hair, Help, Hurting, Ignorance, Illness, Impress, Invention, Investigation, Invisible, Job, Jungle, Kindness, King, Lady, Land, Laugh, Law, Left, Level, Liar, Light, Magic, Marriage, Material, Mathematical, Mechanism, Medication, Money, Mountain, Movement, Movie, Music, Nation, News, Noise, Obligation, Pain, Personality, Philosophy, Picture, Pig, Plan, Plant, Play, Pleasure, Poor, Prison, Professional, Protection, Quality, Reaction, Read, Relationship, Religious, Residence, Road, Sad, Science, Seafood, Sell, Sew, Sexy, Shape, Ship, Show, Sign, Silly, Sin, Skin, Smart, Smiling, Solution, Soul, Sound, Spoke, Star, Student, Stupid, Successful, Sugar, Suspect, Table, Taste, Team, Texture, Time, Tool, Toy, Tree, Trial, Tried, Typical, Unaware, Usable, Useless, Vacation, War, Wash, Weak, Wear, Weather, Willingly, Word.

A.2 fMRI responses and preprocessing

Pereira et al. (2018) preprocessed the fMRI data by estimating the response to each stimulus using a general linear model in which each stimulus presentation (sentence or word and picture) was modeled with a boxcar function convolved with the canonical hemodynamic response; the voxelwise beta estimates are what we referred to as *brain responses* or *vixel-wise activations* throughout the paper.

B Extended Methods

B.1 Model details

We use both VLMs and their language-only counterparts off-the-shelf, as publicly available in the HuggingFace[9] library. Below, we report the HuggingFace IDs of the specific model implementations used in our experiments:

- `openai/clip-vit-large-patch14`
- `kakaobrain/align-base`
- `unc-nlp/lxmert-base-uncased`
- `uclanlp/visualbert-vqa-coco-pre`
- `HuggingFaceM4/idefics2-8b`
- `llava-hf/llama3-llava-next-8b-hf`
- `bert-base-uncased`
- `mistralai/Mistral-7B-v0.1`

[9] https://huggingface.co/

Network	Regions of Interest	
Language	- Posterior temporal lobe - Anterior temporal lobe - Angular gyrus	- Inferior frontal gyrus - Middle frontal gyrus - Orbitan inferior frontal gyrus
Visual	- Parahippocampal place area - Retrosplenial cortex - Transverse occipital sulcus - Lateral occipital area	- Superior temporal sulcus - Fusiform face area - Occipital face area - Extrastriate body area

Table 1: Brain regions of interest (ROIs) corresponding to our investigated functionally localised *language* and *visual* networks.

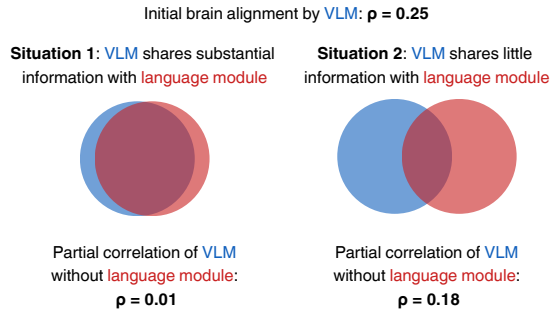

Figure 5: Schematic illustrating situations that can be disambiguated by computing partial correlations. If the initial brain alignment of a VLM is attributable to information substantially shared with the language-only module, the partial correlation will be significantly weaker than the initial correlation.

- `meta-llama/Meta-Llama-3-8B-Instruct`

As for the GloVe representations included in our study, they were extracted using the official vectors[10] pretrained on CommonCrawl (840B tokens version).

B.2 Partial correlation analysis

To better illustrate how partial correlation analysis can help disentangle the representational-alignment contributions by different models, we provide a visualisation in Figure 5.

C Detailed Results

C.1 Comparisons between models

Table 2 provides all the 'raw' (i.e., not yet Bonferroni corrected) p-values we calculate in the main experiment to assess whether pairwise differences between models are statistically significant.

C.2 Layer-wise RSA results

We provide layer-wise RSA results for the sentence condition in Tables 3, 5, 4 and 6. Layer-wise results for the picture condition are reported in Tables 7 and 8. In each of these tables, we boldface the highest correlation value (which was identified before rounding to the second decimal) and check whether it is statistically significantly different from the correlation achieved by each of the other layers by performing pairwise comparisons as described in the main paper. p-values are Bonferroni-corrected, with the number of comparisons amounting to $1 - \#layers$ for each model.

C.3 Picture condition ablation: Additional comparison

To facilitate comparison between VLMs' brain alignment in the picture condition when the image is passed vs. when it is not, we visualise this information in Figure 6.

[10] https://nlp.stanford.edu/projects/glove/

Model Pair	LH Language		Visual	
	Sent.	*Pic.*	*Sent.*	*Pic.*
CLIP - ALIGN	0.001	0.014	0.051	0.0
CLIP - LXMERT	0.0	0.0	0.0	0.229
CLIP - VisualBERT	0.0	0.0	0.0	0.0
CLIP - IDEFICS2	0.745	0.0	0.0	0.0
CLIP - LLaVA	0.0	0.0	0.013	0.0
CLIP - Mistral	0.567	0.0	0.0	0.0
CLIP - Llama3	0.0	0.0	0.555	0.0
CLIP - BERT	0.001	0.0	0.234	0.0
CLIP - GloVe	0.097	0.0	0.001	0.0
ALIGN - LXMERT	0.0	0.004	0.0	0.0
ALIGN - VisualBERT	0.0	0.005	0.0	0.0
ALIGN - IDEFICS2	0.002	0.0	0.0	0.0
ALIGN - LLaVA	0.324	0.0	0.0	0.0
ALIGN - Mistral	0.004	0.0	0.0	0.0
ALIGN - Llama3	0.078	0.0	0.011	0.0
ALIGN - BERT	0.823	0.0	0.447	0.0
ALIGN - GloVe	0.0	0.0	0.147	0.0
LXMERT - VisualBERT	0.155	0.977	0.037	0.0
LXMERT - IDEFICS2	0.0	0.0	0.135	0.0
LXMERT - LLaVA	0.0	0.0	0.033	0.0
LXMERT - Mistral	0.0	0.0	0.982	0.0
LXMERT - Llama3	0.0	0.0	0.0	0.0
LXMERT - BERT	0.0	0.0	0.0	0.0
LXMERT - GloVe	0.008	0.0	0.0	0.0
VisualBERT - IDEFICS2	0.0	0.0	0.552	0.0
VisualBERT - LLaVA	0.0	0.0	0.0	0.0
VisualBERT - Mistral	0.0	0.0	0.035	0.0
VisualBERT - Llama3	0.0	0.0	0.0	0.0
VisualBERT - BERT	0.0	0.0	0.0	0.0
VisualBERT - GloVe	0.0	0.0	0.0	0.0
IDEFICS2 - LLaVA	0.0	0.01	0.0	0.0
IDEFICS2 - Mistral	0.805	0.0	0.129	0.0
IDEFICS2 - Llama3	0.0	0.0	0.0	0.0
IDEFICS2 - BERT	0.004	0.0	0.0	0.0
IDEFICS2 - GloVe	0.047	0.0	0.0	0.063
LLaVA - Mistral	0.0	0.0	0.035	0.0
LLaVA - Llama3	0.438	0.0	0.059	0.0
LLaVA - BERT	0.226	0.0	0.0	0.0
LLaVA - GloVe	0.0	0.0	0.0	0.001
Mistral - Llama3	0.0	0.0	0.0	0.001
Mistral - BERT	0.008	0.0	0.0	0.002
Mistral - GloVe	0.026	0.0	0.0	0.0
Llama3 - BERT	0.047	0.0	0.075	0.819
Llama3 - GloVe	0.0	0.0	0.0	0.0
BERT - GloVe	0.0	0.072	0.027	0.0

Table 2: p-values associated with all pairwise model comparisons in the main experiment before Bonferroni corrections.

Layer	CLIP	ALIGN	LXMERT	VisualBERT	BERT
1	-0.06 (0.0)*	-0.0 (0.58)*	0.07 (0.0)	**0.1** (0.0)	0.01 (0.13)
2	0.05 (0.0)	-0.01 (0.22)*	0.07 (0.0)	0.08 (0.0)	**0.02** (0.0)
3	0.03 (0.0)	0.01 (0.13)	0.07 (0.0)	0.07 (0.0)	0.02 (0.0)
4	0.02 (0.0)*	**0.02** (0.01)	**0.08 (0.0)**	0.07 (0.0)	0.02 (0.01)
5	0.01 (0.43)*	0.01 (0.2)	0.06 (0.0)	0.07 (0.0)	0.02 (0.03)
6	-0.0 (0.64)*	-0.0 (0.9)	0.05 (0.0)*	0.07 (0.0)*	0.02 (0.01)
7	0.01 (0.13)*	0.0 (0.81)	0.04 (0.0)*	0.06 (0.0)*	0.02 (0.05)
8	0.0 (0.68)*	0.0 (0.81)	0.03 (0.0)*	0.06 (0.0)*	0.01 (0.3)
9	0.01 (0.24)*	0.0 (0.63)	0.04 (0.0)*	0.05 (0.0)*	-0.0 (0.58)*
10	-0.0 (0.66)*	0.0 (0.74)	0.04 (0.0)*	0.03 (0.0)*	-0.03 (0.0)*
11	0.01 (0.12)*	-0.0 (0.97)	0.03 (0.0)*	0.01 (0.22)*	-0.05 (0.0)*
12	0.01 (0.1)*	0.0 (0.69)	-0.03 (0.0)*	-0.0 (0.57)*	-0.05 (0.0)*
13	**0.05** (0.0)	0.02 (0.02)	-0.08 (0.0)*	0.0 (0.55)*	-0.08 (0.0)*
14	-	-	0.03 (0.0)*	-	-

Table 3: Layer-wise RSA values (Spearman correlations) for the sentence condition and LH language network, with p-values indicating the significance of the correlation (i.e., whether it is different from 0) in brackets. Note that asterisks indicate whether each correlation is statistically significantly different from the correlation achieved by the best layer (boldfaced value).

Layer	CLIP	ALIGN	LXMERT	VisualBERT	BERT
1	-0.04 (0.0)*	0.0 (0.76)	0.06 (0.0)	**0.09** (0.0)	0.02 (0.01)
2	**0.04** (0.0)	0.0 (1.0)	0.06 (0.0)	0.07 (0.0)	0.02 (0.01)
3	0.03 (0.0)	0.02 (0.05)	0.06 (0.0)	0.07 (0.0)	0.02 (0.02)
4	0.02 (0.0)	**0.02** (0.01)	**0.08** (0.0)	0.08 (0.0)	**0.03** (0.0)
5	0.01 (0.38)*	0.01 (0.47)	0.06 (0.0)	0.08 (0.0)	0.03 (0.0)
6	-0.01 (0.11)*	-0.0 (0.68)*	0.05 (0.0)*	0.07 (0.0)	0.02 (0.01)
7	0.01 (0.07)*	0.0 (0.96)	0.05 (0.0)*	0.08 (0.0)	0.02 (0.01)
8	0.0 (0.85)*	-0.01 (0.33)*	0.04 (0.0)*	0.07 (0.0)	0.02 (0.04)
9	0.02 (0.01)	-0.01 (0.18)*	0.03 (0.0)*	0.05 (0.0)*	-0.0 (0.77)*
10	0.01 (0.48)*	-0.02 (0.02)*	0.04 (0.0)*	0.04 (0.0)*	-0.02 (0.0)*
11	0.02 (0.0)	-0.02 (0.04)*	0.03 (0.0)*	0.02 (0.03)*	-0.03 (0.0)*
12	0.01 (0.1)*	-0.01 (0.08)*	-0.02 (0.01)*	0.02 (0.04)*	-0.02 (0.0)*
13	0.01 (0.5)*	-0.0 (0.66)*	-0.05 (0.0)*	0.03 (0.0)*	-0.04 (0.0)*
14	-	-	0.05 (0.0)*	-	-

Table 4: Layer-wise RSA values (Spearman correlations) for the sentence condition and visual network, with p-values indicating the significance of the correlation (i.e., whether it is different from 0) in brackets. Note that asterisks indicate whether each correlation is statistically significantly different from the correlation achieved by the best layer (boldfaced value).

Layer	LLaVA	IDEFICS2	Llama3	Mistral
1	-0.05 (0.0)*	-0.01 (0.39)*	-0.05 (0.0)*	-0.01 (0.39)*
2	-0.02 (0.03)*	0.02 (0.01)*	-0.03 (0.0)*	0.02 (0.02)*
3	0.0 (0.71)	-0.01 (0.23)*	-0.02 (0.0)*	-0.01 (0.48)*
4	0.01 (0.5)	-0.03 (0.0)*	0.0 (0.99)	-0.05 (0.0)*
5	-0.0 (0.83)	-0.03 (0.0)*	-0.0 (0.91)	-0.04 (0.0)*
6	-0.03 (0.0)*	0.02 (0.05)*	-0.02 (0.03)	0.01 (0.45)*
7	-0.05 (0.0)*	-0.01 (0.29)*	-0.04 (0.0)*	-0.02 (0.02)*
8	-0.05 (0.0)*	-0.03 (0.0)*	-0.04 (0.0)*	-0.02 (0.0)*
9	-0.07 (0.0)*	-0.06 (0.0)*	-0.06 (0.0)*	-0.04 (0.0)*
10	-0.07 (0.0)*	-0.07 (0.0)*	-0.07 (0.0)*	-0.05 (0.0)*
11	-0.08 (0.0)*	-0.06 (0.0)*	-0.08 (0.0)*	-0.05 (0.0)*
12	-0.07 (0.0)*	-0.05 (0.0)*	-0.08 (0.0)*	-0.05 (0.0)*
13	-0.08 (0.0)*	-0.05 (0.0)*	-0.09 (0.0)*	-0.05 (0.0)*
14	-0.09 (0.0)*	-0.06 (0.0)*	-0.09 (0.0)*	-0.06 (0.0)*
15	-0.09 (0.0)*	-0.04 (0.0)*	-0.09 (0.0)*	-0.04 (0.0)*
16	-0.06 (0.0)*	-0.04 (0.0)*	-0.04 (0.0)*	-0.04 (0.0)*
17	-0.05 (0.0)*	-0.02 (0.01)*	-0.01 (0.16)	-0.02 (0.01)*
18	-0.07 (0.0)*	-0.04 (0.0)*	-0.02 (0.01)*	-0.03 (0.0)*
19	-0.04 (0.0)*	-0.03 (0.0)*	0.0 (0.81)	-0.03 (0.0)*
20	-0.04 (0.0)*	0.0 (0.69)*	**0.01** (0.41)	0.02 (0.04)*
21	-0.05 (0.0)*	0.02 (0.0)	0.0 (0.54)	0.04 (0.0)
22	-0.04 (0.0)*	0.03 (0.0)	-0.0 (0.63)	0.04 (0.0)
23	-0.02 (0.02)*	0.04 (0.0)	0.0 (0.6)	**0.04** (0.0)
24	-0.02 (0.0)*	**0.05** (0.0)	-0.01 (0.18)	0.04 (0.0)
25	-0.02 (0.01)*	0.04 (0.0)	-0.02 (0.02)	0.03 (0.0)
26	-0.02 (0.0)*	0.03 (0.0)	-0.03 (0.0)*	0.02 (0.01)
27	-0.03 (0.0)*	0.03 (0.0)	-0.03 (0.0)*	0.02 (0.03)*
28	-0.03 (0.0)*	0.02 (0.0)	-0.03 (0.0)*	0.01 (0.18)*
29	-0.03 (0.0)*	0.01 (0.26)*	-0.04 (0.0)*	-0.0 (0.81)*
30	-0.02 (0.0)*	-0.01 (0.33)*	-0.03 (0.0)*	-0.01 (0.17)*
31	-0.03 (0.0)*	-0.02 (0.05)*	-0.04 (0.0)*	-0.01 (0.36)*
32	-0.02 (0.04)*	-0.02 (0.05)*	-0.04 (0.0)*	-0.01 (0.11)*
33	**0.01** (0.1)	-0.01 (0.52)*	0.0 (0.57)	-0.01 (0.52)*

Table 5: Layer-wise RSA values (Spearman correlations) for the sentence condition and LH language network, with p-values indicating the significance of the correlation (i.e., whether it is different from 0) in brackets. Note that asterisks indicate whether each correlation is statistically significantly different from the correlation achieved by the best layer (boldfaced value).

Layer	LLaVA	IDEFICS2	Llama3	Mistral
1	-0.01 (0.08)*	0.01 (0.27)*	-0.02 (0.05)*	0.01 (0.27)*
2	-0.02 (0.04)*	-0.01 (0.2)*	-0.02 (0.01)*	0.0 (0.75)*
3	0.01 (0.07)*	0.02 (0.0)*	-0.01 (0.47)*	0.02 (0.01)*
4	0.03 (0.0)*	-0.01 (0.28)*	0.0 (0.95)*	-0.02 (0.06)*
5	0.01 (0.11)*	-0.01 (0.08)*	0.0 (0.71)*	-0.02 (0.0)*
6	0.01 (0.07)*	0.0 (0.93)*	0.02 (0.02)	-0.0 (0.86)*
7	-0.0 (0.54)*	-0.0 (0.6)*	-0.0 (0.78)*	-0.01 (0.32)*
8	-0.01 (0.08)*	-0.02 (0.05)*	-0.02 (0.04)*	-0.01 (0.17)*
9	-0.02 (0.0)*	-0.03 (0.0)*	-0.03 (0.0)*	-0.01 (0.1)*
10	-0.03 (0.0)*	-0.05 (0.0)*	-0.04 (0.0)*	-0.03 (0.0)*
11	-0.03 (0.0)*	-0.05 (0.0)*	-0.05 (0.0)*	-0.03 (0.0)*
12	-0.03 (0.0)*	-0.03 (0.0)*	-0.05 (0.0)*	-0.03 (0.0)*
13	-0.03 (0.0)*	-0.03 (0.0)*	-0.05 (0.0)*	-0.03 (0.0)*
14	-0.06 (0.0)*	-0.03 (0.0)*	-0.06 (0.0)*	-0.03 (0.0)*
15	-0.06 (0.0)*	-0.03 (0.0)*	-0.06 (0.0)*	-0.03 (0.0)*
16	-0.02 (0.01)*	-0.04 (0.0)*	-0.01 (0.08)*	-0.04 (0.0)*
17	-0.01 (0.15)*	-0.02 (0.01)*	-0.0 (0.81)*	-0.02 (0.0)*
18	-0.03 (0.0)*	-0.03 (0.0)*	-0.02 (0.05)*	-0.02 (0.0)*
19	0.0 (0.54)*	-0.03 (0.0)*	0.01 (0.29)*	-0.03 (0.0)*
20	0.03 (0.0)*	0.03 (0.0)*	0.02 (0.01)	0.03 (0.0)*
21	0.02 (0.01)*	0.04 (0.0)*	0.02 (0.0)	0.04 (0.0)*
22	0.03 (0.0)	0.06 (0.0)	0.03 (0.0)	0.06 (0.0)
23	0.05 (0.0)	0.08 (0.0)	**0.04** (0.0)	0.07 (0.0)
24	0.06 (0.0)	0.09 (0.0)	0.04 (0.0)	**0.07** (0.0)
25	**0.06** (0.0)	**0.09** (0.0)	0.03 (0.0)	0.07 (0.0)
26	0.05 (0.0)	0.09 (0.0)	0.02 (0.0)	0.07 (0.0)
27	0.05 (0.0)	0.09 (0.0)	0.02 (0.01)	0.07 (0.0)
28	0.05 (0.0)	0.08 (0.0)	0.02 (0.05)*	0.06 (0.0)
29	0.04 (0.0)	0.06 (0.0)	0.01 (0.15)*	0.05 (0.0)*
30	0.05 (0.0)	0.06 (0.0)*	0.01 (0.12)*	0.05 (0.0)
31	0.03 (0.0)*	0.05 (0.0)*	-0.01 (0.36)*	0.05 (0.0)
32	0.02 (0.0)*	0.05 (0.0)*	-0.02 (0.06)*	0.04 (0.0)*
33	0.01 (0.31)*	0.06 (0.0)*	0.0 (0.9)*	0.06 (0.0)

Table 6: Layer-wise RSA values (Spearman correlations) for the sentence condition and visual network, with p-values indicating the significance of the correlation (i.e., whether it is different from 0) in brackets. Note that asterisks indicate whether each correlation is statistically significantly different from the correlation achieved by the best layer (boldfaced value).

Layer	LXMERT	VisualBERT	BERT	LLaVA	IDEFICS2	Llama3	Mistral
1	0.23 (0.0)*	0.33 (0.0)	-0.01 (0.11)*	0.04 (0.0)*	-0.05 (0.0)*	-0.11 (0.0)*	-0.12 (0.0)*
2	0.23 (0.0)*	0.31 (0.0)	**0.11** (0.0)	0.05 (0.0)*	0.06 (0.0)*	-0.1 (0.0)*	-0.1 (0.0)*
3	0.24 (0.0)*	0.32 (0.0)	0.06 (0.0)*	-0.08 (0.0)*	-0.0 (0.56)*	-0.13 (0.0)*	-0.13 (0.0)*
4	0.23 (0.0)*	0.32 (0.0)	0.07 (0.0)*	-0.0 (0.83)*	-0.04 (0.0)*	-0.11 (0.0)*	-0.1 (0.0)*
5	0.23 (0.0)*	0.32 (0.0)	0.01 (0.5)*	0.03 (0.0)*	-0.04 (0.0)*	-0.15 (0.0)*	-0.09 (0.0)*
6	0.26 (0.0)*	0.31 (0.0)	-0.07 (0.0)*	0.07 (0.0)*	-0.03 (0.0)*	-0.13 (0.0)*	-0.05 (0.0)*
7	0.26 (0.0)*	0.3 (0.0)*	-0.08 (0.0)*	0.1 (0.0)*	-0.06 (0.0)*	-0.13 (0.0)*	-0.06 (0.0)*
8	0.28 (0.0)*	0.31 (0.0)	-0.07 (0.0)*	0.09 (0.0)*	-0.04 (0.0)*	-0.13 (0.0)*	-0.04 (0.0)*
9	0.27 (0.0)*	0.31 (0.0)*	-0.04 (0.0)*	0.08 (0.0)*	-0.06 (0.0)*	-0.13 (0.0)*	-0.05 (0.0)*
10	0.27 (0.0)*	0.31 (0.0)*	-0.03 (0.0)*	0.11 (0.0)*	-0.01 (0.08)*	-0.13 (0.0)*	-0.05 (0.0)*
11	0.27 (0.0)*	0.32 (0.0)	-0.02 (0.01)*	0.09 (0.0)*	-0.01 (0.06)*	-0.13 (0.0)*	-0.07 (0.0)*
12	0.28 (0.0)*	**0.33** (0.0)	0.02 (0.05)*	0.07 (0.0)*	-0.03 (0.0)*	-0.13 (0.0)*	-0.08 (0.0)*
13	0.27 (0.0)*	0.33 (0.0)	0.01 (0.16)*	0.04 (0.0)*	-0.04 (0.0)*	-0.13 (0.0)*	-0.09 (0.0)*
14	**0.33** (0.0)	-	-	0.07 (0.0)*	-0.04 (0.0)*	-0.13 (0.0)*	-0.1 (0.0)*
15	-	-	-	0.07 (0.0)*	-0.04 (0.0)*	-0.13 (0.0)*	-0.1 (0.0)*
16	-	-	-	-0.02 (0.03)*	-0.1 (0.0)*	-0.13 (0.0)*	-0.14 (0.0)*
17	-	-	-	-0.02 (0.06)*	-0.08 (0.0)*	-0.14 (0.0)*	-0.16 (0.0)*
18	-	-	-	0.02 (0.04)*	-0.14 (0.0)*	-0.14 (0.0)*	-0.19 (0.0)*
19	-	-	-	-0.02 (0.01)*	-0.1 (0.0)*	-0.13 (0.0)*	-0.17 (0.0)*
20	-	-	-	-0.03 (0.0)*	-0.06 (0.0)*	-0.12 (0.0)*	-0.15 (0.0)*
21	-	-	-	-0.05 (0.0)*	-0.01 (0.08)*	-0.11 (0.0)*	-0.16 (0.0)*
22	-	-	-	-0.04 (0.0)*	0.04 (0.0)*	-0.09 (0.0)*	-0.15 (0.0)*
23	-	-	-	0.03 (0.0)*	0.12 (0.0)*	-0.07 (0.0)	-0.11 (0.0)*
24	-	-	-	0.06 (0.0)*	0.14 (0.0)	-0.07 (0.0)	-0.09 (0.0)*
25	-	-	-	0.09 (0.0)*	0.15 (0.0)	-0.07 (0.0)	-0.06 (0.0)*
26	-	-	-	0.1 (0.0)*	0.16 (0.0)	-0.07 (0.0)	-0.04 (0.0)*
27	-	-	-	0.09 (0.0)*	**0.16** (0.0)	-0.06 (0.0)	-0.02 (0.0)*
28	-	-	-	0.09 (0.0)*	0.15 (0.0)	-0.07 (0.0)	-0.01 (0.07)*
29	-	-	-	0.1 (0.0)*	0.14 (0.0)*	-0.06 (0.0)	-0.01 (0.07)*
30	-	-	-	0.12 (0.0)*	0.14 (0.0)*	-0.07 (0.0)	-0.04 (0.0)*
31	-	-	-	**0.18** (0.0)	0.14 (0.0)	-0.07 (0.0)	-0.04 (0.0)*
32	-	-	-	0.17 (0.0)	0.14 (0.0)*	-0.08 (0.0)*	-0.06 (0.0)*
33	-	-	-	0.1 (0.0)*	0.1 (0.0)*	**-0.05** (0.0)	**0.04** (0.0)

Table 7: Layer-wise RSA values (Spearman correlations) for the picture condition and LH language network, with p-values indicating the significance of the correlation (i.e., whether it is different from 0) in brackets. Note that asterisks indicate whether each correlation is statistically significantly different from the correlation achieved by the best layer (boldfaced value).

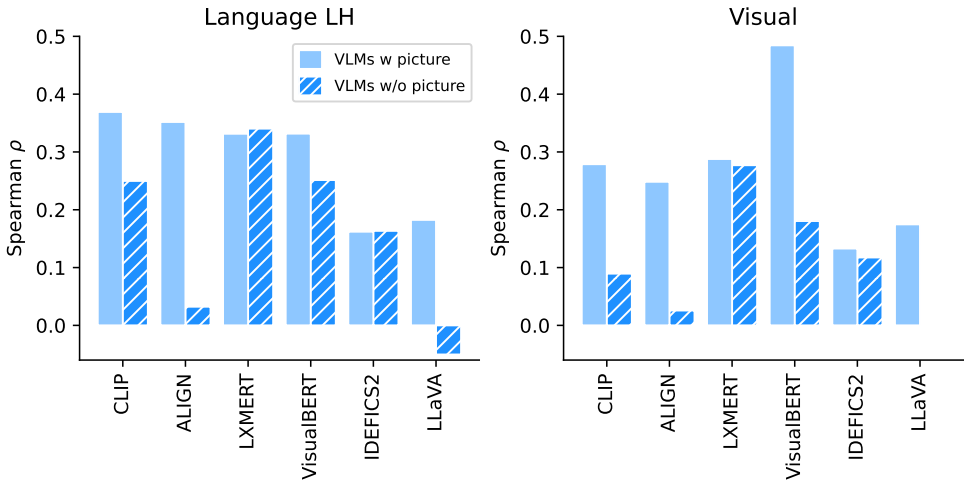

Figure 6: Initial VLM results from RSA analysis in the picture condition vs. results obtained when passing only concept words (without images).

Layer	LXMERT	VisualBERT	BERT	LLaVA	IDEFICS2	Llama3	Mistral
1	0.2 (0.0)*	**0.48** (0.0)	-0.01 (0.36)*	-0.02 (0.02)*	0.01 (0.08)*	-0.07 (0.0)*	-0.04 (0.0)*
2	0.2 (0.0)*	0.47 (0.0)	**0.08** (0.0)	0.04 (0.0)*	0.1 (0.0)*	-0.03 (0.0)*	-0.02 (0.0)*
3	0.21 (0.0)*	0.46 (0.0)*	0.06 (0.0)	-0.0 (0.72)*	0.05 (0.0)*	-0.04 (0.0)*	-0.09 (0.0)*
4	0.2 (0.0)*	0.46 (0.0)*	0.06 (0.0)	0.04 (0.0)*	0.05 (0.0)*	-0.03 (0.0)*	-0.05 (0.0)*
5	0.19 (0.0)*	0.45 (0.0)*	0.02 (0.02)*	0.06 (0.0)*	0.04 (0.0)*	-0.05 (0.0)*	-0.04 (0.0)*
6	0.21 (0.0)*	0.43 (0.0)*	-0.03 (0.0)*	0.08 (0.0)*	0.05 (0.0)*	-0.05 (0.0)*	0.0 (0.72)*
7	0.22 (0.0)*	0.42 (0.0)*	-0.04 (0.0)*	0.11 (0.0)*	0.03 (0.0)*	-0.05 (0.0)*	-0.02 (0.03)*
8	0.22 (0.0)*	0.43 (0.0)*	-0.02 (0.01)*	0.11 (0.0)*	0.02 (0.04)*	-0.05 (0.0)*	0.01 (0.35)*
9	0.22 (0.0)*	0.42 (0.0)*	0.02 (0.03)*	0.11 (0.0)*	0.01 (0.07)*	-0.05 (0.0)*	-0.0 (0.62)*
10	0.22 (0.0)*	0.41 (0.0)*	-0.0 (0.91)*	0.14 (0.0)*	0.03 (0.0)*	-0.05 (0.0)*	-0.01 (0.46)*
11	0.25 (0.0)*	0.4 (0.0)*	-0.01 (0.48)*	0.13 (0.0)*	0.03 (0.0)*	-0.05 (0.0)*	-0.02 (0.01)*
12	0.27 (0.0)	0.39 (0.0)*	0.02 (0.0)*	0.12 (0.0)*	0.02 (0.01)*	-0.05 (0.0)*	-0.02 (0.0)*
13	0.27 (0.0)	0.36 (0.0)*	0.03 (0.0)*	0.08 (0.0)*	0.03 (0.0)*	-0.05 (0.0)*	-0.03 (0.0)*
14	**0.29** (0.0)	-	-	0.1 (0.0)*	0.03 (0.0)*	-0.05 (0.0)*	-0.04 (0.0)*
15	-	-	-	0.08 (0.0)*	0.02 (0.0)*	-0.05 (0.0)*	-0.04 (0.0)*
16	-	-	-	0.04 (0.0)*	-0.01 (0.43)*	-0.05 (0.0)*	-0.06 (0.0)*
17	-	-	-	0.05 (0.0)*	0.0 (0.84)*	-0.05 (0.0)*	-0.06 (0.0)*
18	-	-	-	0.04 (0.0)*	-0.03 (0.0)*	-0.05 (0.0)*	-0.09 (0.0)*
19	-	-	-	0.05 (0.0)*	-0.01 (0.12)*	-0.04 (0.0)*	-0.08 (0.0)*
20	-	-	-	0.02 (0.01)*	0.0 (0.96)*	-0.05 (0.0)*	-0.08 (0.0)*
21	-	-	-	0.02 (0.01)*	0.07 (0.0)*	-0.05 (0.0)*	-0.05 (0.0)*
22	-	-	-	0.04 (0.0)*	0.08 (0.0)*	-0.02 (0.01)*	-0.08 (0.0)*
23	-	-	-	0.07 (0.0)*	0.11 (0.0)	-0.0 (0.6)*	-0.06 (0.0)*
24	-	-	-	0.09 (0.0)*	0.12 (0.0)	-0.0 (0.98)*	-0.05 (0.0)*
25	-	-	-	0.1 (0.0)*	0.13 (0.0)	0.0 (0.63)*	-0.03 (0.0)*
26	-	-	-	0.1 (0.0)*	0.13 (0.0)	0.0 (0.57)*	-0.02 (0.03)*
27	-	-	-	0.11 (0.0)*	**0.13** (0.0)	0.01 (0.29)*	-0.0 (0.67)*
28	-	-	-	0.11 (0.0)*	0.11 (0.0)	0.01 (0.41)*	0.01 (0.4)*
29	-	-	-	0.12 (0.0)*	0.1 (0.0)*	0.01 (0.41)*	0.01 (0.19)*
30	-	-	-	0.13 (0.0)*	0.09 (0.0)*	0.0 (0.59)*	0.0 (0.97)*
31	-	-	-	**0.17** (0.0)	0.11 (0.0)*	0.0 (0.71)*	-0.01 (0.4)*
32	-	-	-	0.17 (0.0)	0.1 (0.0)*	-0.0 (0.67)*	-0.02 (0.05)*
33	-	-	-	0.17 (0.0)	0.08 (0.0)*	**0.08** (0.0)	**0.05** (0.0)

Table 8: Layer-wise RSA values (Spearman correlations) for the picture condition and visual network, with p-values indicating the significance of the correlation (i.e., whether it is different from 0) in brackets. Note that asterisks indicate whether each correlation is statistically significantly different from the correlation achieved by the best layer (boldfaced value).