Experiential Semantic Information and Brain Alignment: Are Multimodal Models Better than Language Models?

Anna Bavaresco, Raquel Fernández
Institute for Logic, Language and Computation
University of Amsterdam
{a.bavaresco, raquel.fernandez}@uva.nl

Abstract

A common assumption in Computational Linguistics is that text representations learnt by multimodal models are richer and more human-like than those by language-only models, as they are grounded in images or audio—similar to how human language is grounded in real-world experiences. However, empirical studies checking whether this is true are largely lacking. We address this gap by comparing word representations from contrastive multimodal models vs. language-only ones in the extent to which they capture experiential information—as defined by an existing norm-based 'experiential model'—and align with human fMRI responses. Our results indicate that, surprisingly, language-only models are superior to multimodal ones in both respects. Additionally, they learn more unique brain-relevant semantic information beyond that shared with the experiential model. Overall, our study highlights the need to develop computational models that better integrate the complementary semantic information provided by multimodal data sources.

https://github.com/dmg-illc/exp-info-models-brain

1 Introduction

How to link language representations to the real-world entities they refer to is a long-standing issue within semantics—the 'symbol-grounding problem' (Harnad, 1990; Bender and Koller, 2020). With the advent of large language models (LLMs) learning astounding linguistic abilities purely from text, this question has been reframed as the 'vector-grounding problem' (Mollo and Millière, 2023), gaining new relevance. While some researchers think that word meanings should be intended as purely symbolic (Fodor, 1983), others believe that words have meanings *precisely because* they are linked to specific entities, experiences or notions (Barsalou, 2008). Supporters of the latter view stress that human language acquisition is situated in a rich multimodal environment, where new words are learnt through interactions with objects and people (Vigliocco et al., 2014). Theories of embodied cognition further highlight the importance of linking words to concrete experience not only for their acquisition but also for their comprehension. Indeed, according to these theories, understanding sentences involves engaging perceptual, motor or emotional simulations of their content (for an overview, see Kaschak et al., 2024).

The idea of obtaining richer semantic representations by learning them from sources other than text, such as images or audio, has inspired a great deal of computational work, from early attempts at concatenating image and text embeddings (e.g., Bruni et al., 2014; Kiela and Bottou, 2014; Derby et al., 2018; Davis et al., 2019) to the most recent large vision-language models (LVLMs, e.g., Li et al., 2023; Wang et al., 2024; Liu et al., 2024; Deitke et al., 2024; Laurençon et al., 2024). Some of these works aimed to obtain language representations aligning more closely with human responses, such as similarity judgments, while others were more oriented towards improving performance on benchmarks or downstream applications. Regardless of the end goal, all these works present multimodality as a *desideratum*, assuming that images provide additional semantic information that cannot be learnt from text alone; however, there is little to no work investigating *which* these semantic aspects are. In this paper, we aim to fill this gap by addressing the following question: *Do multimodal models learn some facets of meaning related to perceptual experiences that language-only models cannot capture?*

Operationalising the 'extra-linguistic' information that multimodal models are allegedly learning is a prerequisite for approaching this issue. We did this by relying on a semantic model introduced by Fernandino et al. (2022) to capture 'experiential information'. This cognitive model represents words as n-dimensional arrays where each entry

Proceedings of the 29th Conference on Computational Natural Language Learning, pages 141–155

corresponds to aggregated human ratings on a predefined experiential attribute (e.g., *Vision*, *Motion*, *Harm*). We focused on a set of nouns and evaluated the alignment between their representations provided by the experiential model and those by comparable unimodal (language-only) and multimodal (vision-language and audio-language) computational models. This analysis allowed us to uncover if multimodal models indeed reflect more semantic information than language-only models. Next, we checked whether capturing experiential information translates into higher alignment with brain responses recorded with functional magnetic resonance imaging (fMRI) to the same set of nouns.

Our findings indicate several interesting trends. First, both vision-language and language-only models exhibit significant alignment with the experiential model and brain responses, while the audio-language model displays weak or non-significant correlations. Second, this alignment is more pronounced for language-only models, which appear to capture a great deal of brain-relevant information beyond experiential. Lastly, language-only models remain more brain-aligned than vision-language models even when focusing on a set of more concrete words, although the gap is reduced. Overall, our study shows that current multimodal models learn *less* brain-relevant information—both experiential and beyond—than comparable language-only models, highlighting the need to explore different approaches to construct multimodal word representations.

2 Background

2.1 Embodied cognition

Embodied cognition identifies a suite of theoretical frameworks holding that language is understood by perceptual, emotional, or motor simulations of its content (e.g., Barsalou, 1999; Glenberg and Gallese, 2012; Zwaan, 2014; Pulvermüller, 2018). This general principle has received empirical support from multiple studies, both behavioural and neuroscientific.

For example, a series of works on the Action-sentence Compatibility Effect (ACE, Glenberg and Kaschak, 2002) and its subsequent variants (Borreggine and Kaschak, 2006; Zwaan and Taylor, 2006; Bub and Masson, 2012) revealed a significant difference in reaction times—attributed to motor simulations—when participants had to respond to a sentence (e.g., *You passed the note to Art*) with a movement matching (extending their arm) vs. non-matching (retreating their arm) that mentioned in the sentence. Similarly, the sentence-picture verification task (Stanfield and Zwaan, 2001), where participants have to respond to a picture that is either compatible (an eagle with its wings outstretched) or incompatible (an eagle with its wings folded) with a sentence (*The eagle is in the sky*), and its variations (Connell, 2007; Hoeben Mannaert et al., 2017) have also been widely used to demonstrate the occurrence of perceptual simulation during language comprehension. In parallel, a line of neuroscientific studies have found evidence that semantic processing may activate motor (among others, Hauk et al., 2004; Tettamanti et al., 2005; Aziz-Zadeh et al., 2006) and perceptual brain regions (Kiefer et al., 2008; Van Dam et al., 2012).

2.2 Multimodal models of semantics

Embodied cognition and related ideas, such as *visual grounding*, have percolated from Cognitive Science to Computational Linguistics, motivating attempts to build semantic models that learn representations from data sources beyond text. Early efforts in this direction (e.g., Bruni et al., 2014; Kiela and Bottou, 2014; Lazaridou et al., 2015; Silberer and Lapata, 2012, 2014) were characterised by 1) a focus on developing human-aligned computational models of meaning and 2) limited computational modelling resources (large datasets of paired image-text inputs did not exist at the time, nor did large transformer-based architectures).

Recently, multimodal models have become more powerful and found application on a variety of downstream tasks (e.g., image captioning, image retrieval, or visual question answering). Some seminal works used a contrastive objective to learn aligned image and text representations (Radford et al., 2021; Jia et al., 2021), while others—often inspired by BERT's (Devlin et al., 2019) successes in language modelling—applied its underlying intuitions to the vision-language domain (Tan and Bansal, 2019; Li et al., 2019; Lu et al., 2019; Chen et al., 2020). Finally, state-of-the-art large vision-language models (LVLMs, e.g., Li et al., 2023; Wang et al., 2024; Liu et al., 2024; Deitke et al., 2024; Laurençon et al., 2024), usually combining a large language model (LLM) with an image encoder, can engage in strikingly human-like conversations about images. In contrast to the early attempts at multimodal modelling, these works share 1) a focus on solving, or improving performance

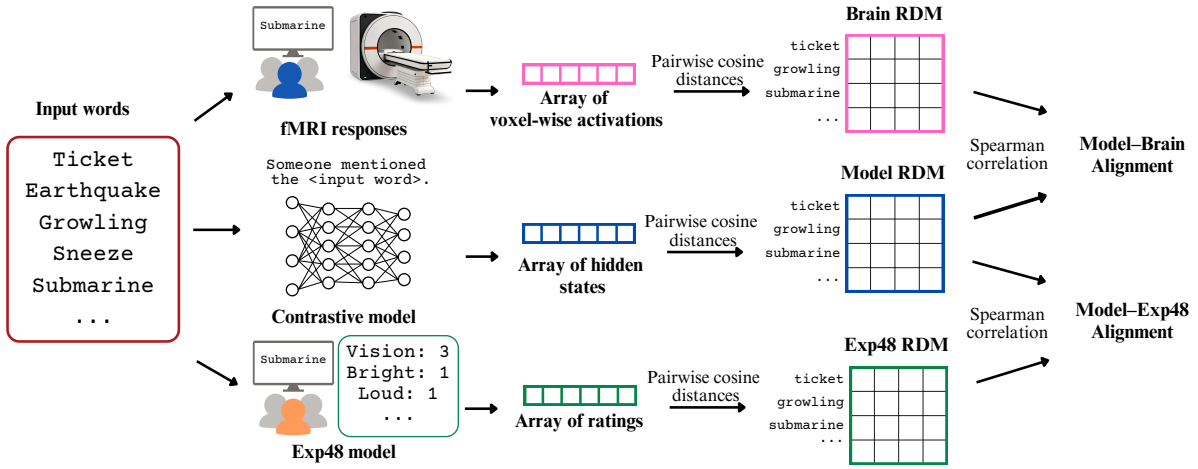

Figure 1: Overview of our experimental setup. Representations for the word stimuli are derived from three different sources: 1) fMRI responses; 2) multimodal and language-only contrastive models; 3) human ratings along the experiential dimensions of the EXP48 model. Next, pairwise distances between these word representations are used to populate representational dissimilarity matrices (RDMs). Finally, alignment between representational spaces is computed by correlating the off-diagonal elements of the RDMs.

on, downstream tasks, and 2) the availability of massive datasets and large models with billions of parameters.

For our experiments, we aimed to leverage models that are powerful while, at the same time, suitable for drawing cognitively-meaningful comparisons. The need to satisfy both constraints prevented us from evaluating state-of-the-art LVLMs; we elaborate more on our model choices in Section 3.2.

2.3 Experiential models of semantics

Recently, a few approaches motivated by embodied cognition have introduced models of semantics aimed at capturing 'experiential information', i.e., aspects of meaning related to how humans ground language in experiences. These experiential models were constructed by asking human annotators to rank words on a set of pre-defined dimensions. For example, Fernandino et al. (2022) introduced an experiential model based on 48 dimensions spanning perceptual, emotional, and action-related constructs. In two fMRI studies, they found that the experiential model yields more brain-aligned word representations than taxonomic and distributional models; additionally, it contributes unique semantic information not represented by the other models.

Similarly, Carota et al. (2024) experimented with a different experiential model based on 11 dimensions and compared its brain alignment against that of a distributional model. Their study revealed significant correlations with brain responses in more ROIs (regions of interest) for the experiential model than for the distributional model. However, an integrative model combining both displayed significant correlations in an even larger number of ROIs, suggesting that experiential and distributional are complementary aspects of human semantic processing.

Despite their merits, experiential models are bounded in their accuracy by an *a priori* selection of dimensions and, relying on human annotations, remain expensive to construct. These limitations open the intriguing question of whether experiential information can be captured by computational models learning semantic representations in a data-driven fashion.

3 Methods

A schematic of our experimental pipeline is provided in Figure 1. In the following, we describe in detail the word stimuli, brain responses, computational models and evaluation procedures.

3.1 Data and experiential model

For our experiments, we used word stimuli, fMRI responses and experiential model from Study 2 by Fernandino et al. (2022).[1] We describe each below.

Word stimuli Word stimuli comprise 320 nouns, half of which refer to *objects* and the other half

[1]These materials have been made publicly available by Fernandino et al. The full list of words and the experiential features can be found at https://www.pnas.org/doi/10.1073/pnas.2108091119#supplementary-materials; fMRI data are available at https://osf.io/87chb/.

to *events*. The 160 object nouns include an equal number of words (40) from four categories (*food*, *vehicles*, *animals*, *tools*); likewise, the event nouns span four semantic subcategories (*social event*, *negative event*, *sound*, *communication*).

fMRI responses fMRI responses were collected from 36 participants. While viewing the above-mentioned word stimuli one at a time, they were instructed to rate the frequency with which they experienced their corresponding entities in daily life. Voxel-wise activations (beta maps) for each noun relative to the mean signal across other nouns were estimated using linear regressions (for additional details, see Fernandino et al., 2022). Here, we focus on the betas from voxels within a 'semantic network ROI' defined by Binder et al. (2009) based on a meta-analysis. Voxel-wise beta coefficients can be arranged in vectors representing the brain response elicited by each noun.

Experiential model The experiential model, hereafter abbreviated as EXP48, represents each word as a set of ratings on 48 pre-defined dimensions capturing different aspects of people's experience with objects/events, e.g., *Vision*, *Hand action* or *Unpleasant*. The ratings were introduced by Binder et al. (2016) as part of a wider set of experiential salience norms; they range from 0 to 6 and were provided by 1743 unique crowdworkers.

3.2 Computational models

Our model choices were motivated by the goal to maximise comparability across architectures. More concretely, we selected three models (language-only, vision-language, and audio-language) comparable in terms of fine-tuning objective—the contrastive one—and architecture—they all have a pretrained BERT (Devlin et al., 2019) as language encoder.[2] One aspect in which these architectures differ is the amount of training data; however, we believe this issue does not invalidate our results and further discuss it in Section 6.

SimCSE (**Sim**ple **C**ontrastive Learning of **S**entence **E**mbeddings, Gao et al., 2021) is a language-only sentence encoder fine-tuned contrastively on 1M sentences randomly sampled from English Wikipedia. Matching pairs for the contrastive objective were created by applying different dropout masks to the same sentence.

MCSE (**M**ultimodal **C**ontrastive Learning of **S**entence **E**mbeddings, Zhang et al., 2022) is a vision-language sentence encoder fine-tuned by jointly optimising a SimCSE objective and a CLIP-like (Radford et al., 2021) objective. The fine-tuning data for the first objective is the same as SimCSE's; as for the CLIP-like objective, where a matching pair was defined by an image and its caption, the fine-tuning data consists of 83K images from MS-COCO (Lin et al., 2014) annotated with multiple captions.

CLAP (**C**ontrastive **L**anguage **A**udio **P**retraining, Wu et al., 2023) is an audio-language model whose language encoder was initialised with pre-trained BERT weights and fine-tuned on audio-caption pairs with a CLIP-like objective. The fine-tuning data includes 633,526 audio-text pairs, with audio clips representing human activities, natural sounds, and audio effects.

For reference, we also tested BERT and VisualBERT (Li et al., 2019) as its visual counterpart.

BERT (Devlin et al., 2019) is a transformer-based language-only model pretrained with two objectives: masked language modelling and next sentence prediction. Its pretraining data includes the BooksCorpus (800M words, Zhu et al., 2015) and English Wikipedia (2500 words). As mentioned above, SimCSE, MCSE and CLAP fine-tuned pretrained BERT architectures.

VisualBERT (Li et al., 2019) is a vision-language model consisting of a BERT-based language encoder (initialised with parameters from pretrained BERT) and a pretrained visual feature extractor based on Faster RCNN (Ren et al., 2015). Its training objectives, which mirror BERT's, were masked language modelling with image input and sentence-image prediction. The vision-language pretraining data comprises MS-COCO and VQA 2.0 (Goyal et al., 2017). Note that this is *not* a contrastive model; we included it for reference as it can be considered as a vision-language extension of BERT, but it is not directly comparable with MCSE, SimCSE and CLAP.

3.3 Extracting representations

Given that all the models we considered were trained to learn contextualised representations

[2] All three models were released with both BERT-based (Devlin et al., 2019) and RoBERTa-based (Liu et al., 2019) implementations. We used the former in all our experiments.

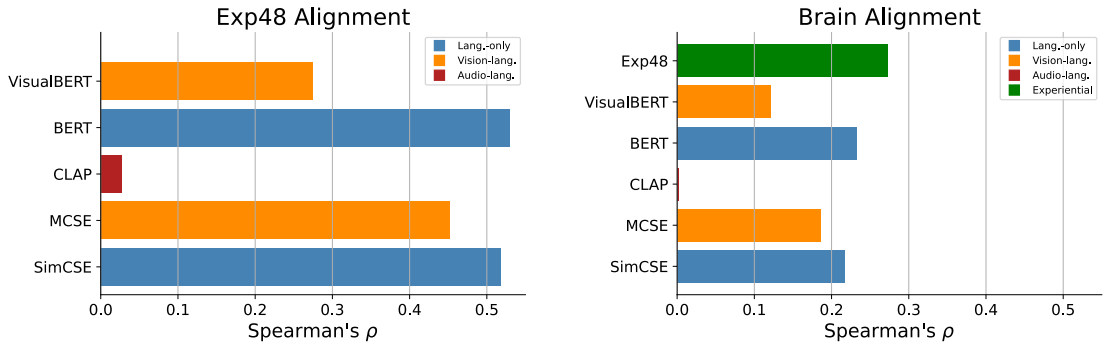

Figure 2: Results from representational similarity analysis. On the left, Spearman correlations quantifying the alignment between word representations from EXP48 and by computational models. On the right, Spearman correlations indicating the alignment between fMRI responses from human participants and word representations by computational models.

from sentences, single words may be an out-of-distribution input. Therefore, following an approach similar to May et al. (2019), we embedded the noun stimuli in a set of generic template sentences (e.g., `Someone mentioned the <word>`, see Appendix A for the complete list) when passing them to the models.[3] For all templates, we derived word representations from the hidden states of each layer; more specifically, we selected the hidden states corresponding to the tokens of the target word and averaged them across templates.

3.4 Alignment evaluation

RSA To compare model representations against EXP48 and brain responses, we used representational similarity analysis (RSA, Kriegeskorte et al., 2008), which quantifies the alignment between two representational spaces (either by two models or by a model and brain responses) as the correlation between representational dissimilarity matrices (RDMs). In our experiments, RDMs were populated with pairwise cosine distances between model representations or fMRI responses for all the unique word-pairs. fMRI RDMs for individual participants were averaged into one aggregated RDM. The alignment between this fMRI RDM and morel-derived RDMs was calculated as a Spearman correlation.

Partial correlations While RSA allows comparing models' alignment with EXP48 or brain responses, it does not reveal whether models explain shared variance or provide independent contributions. Fernandino et al. (2022) computed partial correlations to check how much brain-relevant information EXP48 shared with the other models they considered, i.e., two distributional models (Word2vec and GloVe; Mikolov et al., 2013; Pennington et al., 2014) and two taxonomic models (a WordNet-based model and a categorical one). We used the same approach to determine how much brain-relevant information our tested models share with EXP48 and with each other. Formally, partial correlations can be defined as follows: Consider the RDM from Model A y, the RDM from Model B x, and the RDM of the brain responses z. The partial correlation of Model A without Model B is $\rho(r_i, z_i)$, where $r_i = y_i - \hat{y}_i$ are the residuals from the linear regression with equation $\hat{y}_i = a + bx_i$.

4 Results

4.1 EXP48 and brain alignment across models

We performed RSA to obtain a first measure of model representations' alignment with EXP48 and fMRI responses. This analysis was conducted on model representations averaged across the three layers yielding the highest alignment individually; note that these layers may differ when considering alignment to brain responses vs. EXP48 (see Appendix B.2 for a visualisation of layer-wise alignment). The results from RSA against brain responses and EXP48 are displayed in Figure 2. All Spearman correlations are statistically significant ($p < 0.05$), except for CLAP's correlation with brain responses ($p = 0.70$); we additionally verified that all the pairwise differences between

[3] We empirically verified that passing words within templates rather than in isolation yields higher alignment with both the experiential model and brain responses (see Appendix B.1).

correlations are statistically significant.[4]

An inspection of correlations against EXP48 indicates BERT as the most aligned model ($\rho = 0.53$); SimCSE and MCSE also display moderate correlations with EXP48 ($\rho = 0.52$ and $\rho = 0.45$, respectively). In contrast, CLAP's representations are poorly aligned with EXP48, exhibiting a correlation of just 0.03. A comparison between vision-language models (MCSE and VisualBERT) and their unimodal counterparts (SimCSE and BERT) reveals that the former, surprisingly, reflect less experiential information than the latter.

Regarding alignment with brain responses in the semantic ROI, BERT is again the best model ($\rho = 0.23$), although it remains less brain-aligned than EXP48 ($\rho = 0.27$). All the other models display positive correlations, with the exception of CLAP, whose correlation is not statistically significant ($\rho = 0.00$, $p = 0.70$). Similarly to the EXP48-alignment results, here we found that the language-only models BERT and SimCSE are more brain-aligned than their vision-language extensions VisualBERT and MCSE. We delve deeper into the robustness of this finding in Section 5.

An interesting trend common across results from both RSAs (against EXP48 and fMRI responses) is that representations by SimCSE and MCSE—which have been shown to outperform BERT on semantic text similarity tasks (Gao et al., 2021; Zhang et al., 2022)—are *less* aligned than those by BERT. A potential explanation for this may be that we considered *single-word* representations. Since contrastive fine-tuning, as applied to SimCSE and MCSE, optimises *sentence*-level representations as opposed to *token*-level ones, it could be that some token-level semantic properties initially learnt by BERT got somehow diluted through this process.

4.2 Experiential information vs. unique contribution in models' brain alignment

Results from the partial correlation analysis are displayed in Figure 3, whose left-hand panel shows how much EXP48 representations align with brain responses without the information they share with each of the other models. An interesting observation is that the lowest correlations were obtained when regressing out BERT and SimCSE. This provides an interesting complement to the findings from RSA against EXP48 representations: RSA shows that BERT and SimCSE share substantial representational information with EXP48, and partial correlations suggest that this information is also brain-relevant. Regarding models' brain alignment without EXP48, displayed in Figure 3's right-hand panel, a noteworthy finding is that BERT's and SimCSE's representations are the most brain-aligned even after regressing out EXP48. This suggests that these models learnt some semantic information that is not captured by EXP48 but is still reflected in brain responses.

Additionally, for each model we checked which proportion of its initial brain alignment is attributable to unique contribution as opposed to information shared with EXP48. This can be visualised by comparing the dark-shade bars against the light-shade ones in the right-hand panel of Figure 3. An interesting result revealed by this comparison is that, although MCSE is more brain-aligned than VisualBERT, their unique contribution without EXP48 is the same in absolute value ($\rho = 0.06$); in other terms, 50% of VisualBERT's brain alignment is due to unique information, while in MCSE it is 32%. Regarding BERT and SimCSE, the majority of their initial brain alignment is eroded when regressing out EXP48; however, the asymmetry is not substantial, and the unique contribution accounts for more than 40% of the initial brain alignment in both models. As for CLAP, it exhibits a weak negative correlation that is not statistically significant, confirming that the model does not contribute any brain-relevant information.

Finally, we used partial correlations to compare vision-language models (VLMs) against their language-only counterparts (LMs). We found that neither MCSE ($\rho = 0.00$; $p = 0.60$) nor VisualBERT ($\rho = 0.00$; $p = 0.66$) exhibit statistically significant correlations with brain responses once SimCSE and BERT, respectively, are regressed out. Crucially, this indicates that VLMs did not learn any additional brain-relevant information besides that already captured by their LM counterparts.

5 Assessing Results' Robustness

RSA results revealed a consistent advantage of language-only models over the multimodal ones. This finding contrasts with the expectation—shared across a great deal of work on multimodality and

[4] Statistical significance was determined by applying a Fisher transformation to the correlation coefficients from each pair of models and calculating the p-value associated with the difference between the two z-scores. All p-values were Bonferroni-corrected with $\alpha = 0.05$. The same approach for verifying statistical significance was applied to all correlation comparisons throughout the paper.

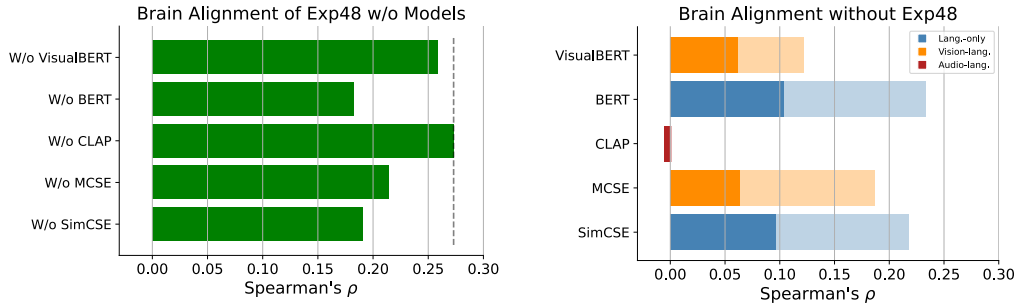

Figure 3: Results from partial correlation analyses. On the left, Spearman correlations between brain responses and the residuals obtained regressing model RDMs out of the EXP48 RDM. The dotted line indicates EXP48's initial brain alignment without removing any information. On the right, Spearman correlations between brain responses and the residuals obtained regressing the EXP48 RDM out of model RDMs. The bars in lighter shades indicate models' initial brain alignment.

language modelling—that training models on diverse data modalities, as opposed to text alone, should yield more human-like language representations. In the following, we present two analyses aimed at assessing the robustness of these findings. Given that the audio-language model CLAP did not achieve a statistically significant brain alignment, we excluded it from further analyses and focused on the remaining vision-language and language-only architectures.

5.1 Do caption-like templates result in improved brain alignment?

As pointed out by Tan and Bansal (2020), image captions are examples of *grounded language*, which differs from other types of natural language along many dimensions. Since the VLMs we evaluated were trained on image-caption pairs, they may have over-fitted to the language present in captions. Therefore, it is possible that the sentence templates we used to obtain contextualised word representations from the models are somehow out-of-distribution for VLMs.

To control for this potential confound, we re-extracted word representations employing different templates, whose structure was modelled around captions (e.g., `There is an <object> in a <place>`, or `A <person> is <verb in -ing> in a <place>`). These structures were identified based on a manual inspection of captions from MS-COCO, which was part of both MCSE's and VisualBERT's training. Given the challenges of creating caption-like templates providing a fitting context for all the word stimuli, we used different sets of templates for each sub-category of words described in Section 3.1 (e.g., `There is a <food-word> on a table in a restaurant` or `A few people gathered for a <social event-word>`). We provide the complete list of templates in Appendix A.

The procedure for calculating brain alignment was the same as that employed in the main experiment. Spearman correlations between model-derived RDMs and the fMRI-derived RDM are displayed in Figure 4. All correlations are statistically significant, as well as correlation differences between models. A comparison across models confirms the trend from the main experiment: Language-only models are more brain-aligned than their vision-language counterparts. This suggests that the finding is robust and not a by-product of the templates where word stimuli were embedded.

The dotted lines in Figure 4 allow comparing the brain alignment model representations achieve when using caption-like templates vs. when using the templates from the main experiment. This comparison reveals that all models—not only VLMs—exhibit higher brain alignment when using caption-like templates. We interpret this as indicating that caption-like templates are not more in-distribution for VLMs, but rather provide a better-specified context that is beneficial to all models.

5.2 Do VLMs yield more brain-aligned representations for objects vs. events?

Provided that VLMs learn additional semantic information, it could be that not all word representations benefit from multimodal training to the same extent; instead, a potential advantage may be more prominent for words referring to visual contents.

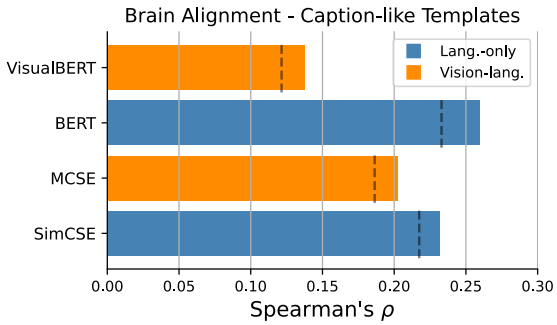

Figure 4: Spearman correlations indicating alignment between model representations extracted using caption-like templates and fMRI responses. Dotted lines indicate the initial correlations obtained with the templates from the main experiment.

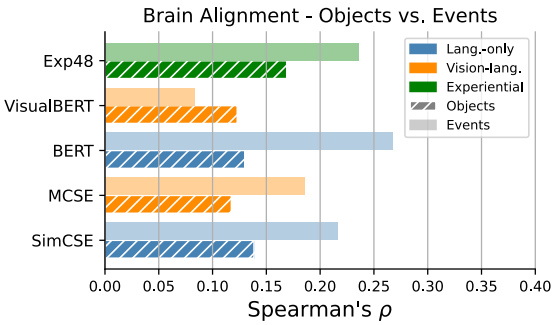

Figure 5: Spearman correlations indicating alignment between model representations and fMRI responses. Correlations are displayed separately for object-words and event-related words.

The words used in our main experiments include nouns from multiple semantic categories (see Section 3.1 for more details), which may largely vary in their degree of 'visual-ness'. Therefore, it is possible that we did not detect additional brain-relevant information learnt by VLMs because we focused on the 'wrong' words.

To check whether this is the case, we further analysed two word subsets with different levels of concreteness. The subsets were identified by leveraging the semantic labels already present in our word set, i.e., *objects* and *events*.[5] We repeated RSA separately for these two word subsets following the same procedure employed in the main experiment.

The results of this analysis are displayed in Figure 5. A first observation is that—for all models except VisualBERT—correlations are statistically significantly stronger for events than objects. This pattern was also reported by Fernandino et al. (2022), who attributed it to "higher variability of pairwise similarities for the neural representations of event concepts".

A second interesting result is that the model ranking we observed analysing the entire word set (BERT > SimCSE > MCSE > VisualBERT) is replicated for events but not for objects, where none of the differences between model correlations is statistically significant. While there is a negative effect overall, further training BERT on image-text pairs (as in VisualBERT) or fine-tuning it with a contrastive objective (as in SimCSE and MCSE) does not significantly alter the initial brain alignment of its object-word representations. Interestingly, EXP48, which we included for reference, is outperformed by BERT on events; however, it remains statistically significantly more brain-aligned than the other models on objects.

Finally, comparing vision-language models against their language-only counterparts shows that BERT and VisualBERT do not significantly differ regarding the brain alignment of their object-word representations, while SimCSE and MCSE do (with SimCSE remaining more aligned).[6] As for event-word representations, SimCSE and BERT are, respectively, significantly more brain-aligned than MCSE and VisualBERT. These results further support the robustness of our initial finding, i.e., that LMs models are more aligned than their VL counterparts. However, the reduced gap between the two model types when considering object-word representations vs. event-word ones suggests that VLMs *do*, comparatively, learn more brain-aligned representations for objects than events.

6 Discussion

While multimodal models are often expected to incorporate additional semantic aspects that language-only models cannot learn, our results reveal that their word representations are *less* aligned with EXP48 and fMRI responses than those by LMs. Moreover, within multimodal models, the vision-language ones show moderate positive correlations with EXP48 and fMRI responses, while the

[5] In their supplementary materials, Fernandino et al. (2022) report that the average concreteness score for *objects* is 4.9, while for *events* it is 3.6.

[6] Note that, since we used Bonferroni corrections, this difference is statistically significant here—but not when comparing all five models—due to a change in the number of relevant comparisons (2 vs. 5).

audio-language one correlates weakly with EXP48 and does not yield a significant correlation with brain responses. Below, we discuss factors that may have played a role in these partially unexpected results.

Amount of training data While being comparable in terms of learning objectives and architecture, SimCSE, MCSE and CLAP still differ in the amount of fine-tuning data. For the SimCSE–MCSE comparison, this does not appear to be a confound: Despite being fine-tuned on *less* sentences than MCSE, SimCSE still proves to be *more* EXP48- and brain-aligned. A potential reason for this may be that the grounded language employed in image captions causes a shift of semantic representations towards more concrete meanings. As for CLAP, the smaller amount of fine-tuning audio-caption pairs, together with its optimising only a CLIP-like objective (without a SimCSE-like one), may have played a role in its poor alignment.

Multimodal pretraining vs. fine-tuning A potential explanation for the inferior performance of multimodal models could be that training on multimodal pairs is not as effective during fine-tuning as it is during pre-training. However, we verified that not even the language encoder from the powerful CLIP (Radford et al., 2021)[7]—pretrained contrastively on 400M image-text pairs—yields more brain-aligned word representations than BERT and SimCSE (see Appendix B.3).

Models from present vs. past studies An interesting result was that the correlations with fMRI responses we observed for SimCSE, MCSE and BERT are higher than those achieved by the computational models (GloVe and Word2vec) tested by Fernandino et al. 2022 (see Appendix B.3). This finding aligns with previous work showing that transformer-based architectures are more predictive of brain responses during language processing than word-level embedding models and recurrent neural networks (Schrimpf et al., 2021). In addition, we found that the LMs and, to a larger extent, the VLMs we tested learn brain-relevant semantic information beyond that captured by EXP48. This partially echoes the results by Carota et al. (2024), with the difference that the computational model included in their study was strictly distributional.

[7]This model was excluded from the main experiment as it is not directly comparable with the other architectures.

Information captured by EXP48 While the ability of EXP48 to model brain responses has been validated by previous research, it may still be a suboptimal model of perceptual experience for two main reasons. First, all dimensions in EXP48, including the more perceptual ones like *Colour* or *Sound*, are somewhat abstract; in this sense, they may fail to capture low-level perceptual information relevant for modelling human word representations and, perhaps, learnt by multimodal models. Second, EXP48 encodes experiential dimensions, e.g., *Pleasant* or *Time*, which are not strictly perceptual and may be hard, if not impossible, to learn for vision-language and audio-language models.

Type of stimuli Our study focuses on single words that are not included in longer text passages. To some extent, our results suggest that this may affect *machine* language processing; indeed, we found that embedding words in sentences, as opposed to passing them to the models as is, yields more brain-aligned representations (see also Appendix B.1). In a similar vein, the amount of context may influence *human* language processing: As observed by Zwaan (2014), context determines the perceptual detail of the mental simulations people engage during language comprehension. Therefore, it may be that the nouns used in the fMRI experiment did not prompt multimodal semantic knowledge enough for it to be detected in our study.

7 Conclusions

Our study provides an in-depth comparison between multimodal and language-only architectures in their ability to capture experiential semantic information and alignment with brain responses. Contrary to common assumptions, we found multimodal models to produce word representations less brain-aligned and experience-informed than language-only models.

These results have several implications for future work. First, they invite caution against assuming that technical innovations allowing models to solve additional downstream tasks should necessarily make them more 'human-like'. Second, they indicate that there is significant room for improving current computational language models so that they learn the brain-relevant experiential information they currently lack—how to concretely achieve this remains an open question.

Limitations

Our experimental setup focuses exclusively on contrastive models which are not state-of-the-art for both linguistic and multimodal downstream tasks. More recent architectures pretrained autoregressively—e.g., models from the LLaVA family (Liu et al., 2024), Molmo (Deitke et al., 2024), or Qwen2.5-VL(Bai et al., 2025)—may exhibit different patterns. However, the complexity of their pre-training and fine-tuning steps makes it hard to set up a controlled comparison ruling out factors such as the amount of training data or training objectives. We therefore explicitly decided to not include this type of model in our investigation. This decision was further informed by preliminary evidence that generative vision-language models achieving stronger performance on downstream tasks are less brain-aligned than previous architectures (Bavaresco et al., 2024).

Acknowledgments

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

This project was funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 819455).

References

Lisa Aziz-Zadeh, Stephen M Wilson, Giacomo Rizzolatti, and Marco Iacoboni. 2006. Congruent embodied representations for visually presented actions and linguistic phrases describing actions. *Current biology*, 16(18):1818–1823.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 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.

Lawrence W Barsalou. 2008. Grounded cognition. *Annu. Rev. Psychol.*, 59(1):617–645.

Anna Bavaresco, Marianne de Heer Kloots, Sandro Pezzelle, and Raquel Fernández. 2024. Modelling multimodal integration in human concept processing with vision-and-language models. *arXiv preprint arXiv:2407.17914*.

Emily M Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5185–5198.

Jeffrey R Binder, Lisa L Conant, Colin J Humphries, Leonardo Fernandino, Stephen B Simons, Mario Aguilar, and Rutvik H Desai. 2016. Toward a brain-based componential semantic representation. *Cognitive neuropsychology*, 33(3-4):130–174.

Jeffrey R Binder, Rutvik H Desai, William W Graves, and Lisa L Conant. 2009. Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral cortex*, 19(12):2767–2796.

Kristin L Borreggine and Michael P Kaschak. 2006. The action–sentence compatibility effect: It's all in the timing. *Cognitive Science*, 30(6):1097–1112.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal Distributional Semantics. *Journal of artificial intelligence research*, 49:1–47.

Daniel N Bub and Michael EJ Masson. 2012. On the dynamics of action representations evoked by names of manipulable objects. *Journal of Experimental Psychology: General*, 141(3):502.

Francesca Carota, Hamed Nili, Nikolaus Kriegeskorte, and Friedemann Pulvermüller. 2024. Experientially-grounded and distributional semantic vectors uncover dissociable representations of conceptual categories. *Language, Cognition and Neuroscience*, 39(8):1020–1044.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: UNiversal Image-TExt Representation Learning. In *European Conference on Computer Vision*, pages 104–120.

Louise Connell. 2007. Representing object colour in language comprehension. *Cognition*, 102(3):476–485.

Christopher Davis, Luana Bulat, Anita Lilla Vero, and Ekaterina Shutova. 2019. Deconstructing multimodality: visual properties and visual context in human semantic processing. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 118–124, Minneapolis, Minnesota. Association for Computational Linguistics.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. 2024. Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Multimodal Models. *CoRR*.

Steven Derby, Paul Miller, Brian Murphy, and Barry Devereux. 2018. Using Sparse Semantic Embeddings Learned from Multimodal Text and Image Data to Model Human Conceptual Knowledge. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 260–270, Brussels, Belgium. Association for Computational Linguistics.

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.

Leonardo Fernandino, Jia-Qing Tong, Lisa L Conant, Colin J Humphries, and Jeffrey R Binder. 2022. Decoding the information structure underlying the neural representation of concepts. *Proceedings of the National Academy of Sciences*, 119(6):e2108091119.

Jerry A Fodor. 1983. *The modularity of mind*. MIT press.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Arthur M Glenberg and Vittorio Gallese. 2012. Action-based language: A theory of language acquisition, comprehension, and production. *cortex*, 48(7):905–922.

Arthur M Glenberg and Michael P Kaschak. 2002. Grounding language in action. *Psychonomic bulletin & review*, 9(3):558–565.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Olaf Hauk, Ingrid Johnsrude, and Friedemann Pulvermüller. 2004. Somatotopic representation of action words in human motor and premotor cortex. *Neuron*, 41(2):301–307.

Lara N Hoeben Mannaert, Katinka Dijkstra, and Rolf A Zwaan. 2017. Is color an integral part of a rich mental simulation? *Memory & cognition*, 45:974–982.

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.

Michael P Kaschak, Michael Long, and Julie Madden. 2024. Embodied Approaches to Language Comprehension. In *The Routledge Handbook of Embodied Cognition*, pages 191–199. Routledge.

Markus Kiefer, Eun-Jin Sim, Bärbel Herrnberger, Jo Grothe, and Klaus Hoenig. 2008. The sound of concepts: Four markers for a link between auditory and conceptual brain systems. *Journal of Neuroscience*, 28(47):12224–12230.

Douwe Kiela and Léon Bottou. 2014. Learning Image Embeddings using Convolutional Neural Networks for Improved Multi-Modal Semantics. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

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? In *Advances in Neural Information Processing Systems*, volume 37, pages 87874–87907. Curran Associates, Inc.

Angeliki Lazaridou, Marco Baroni, et al. 2015. Combining Language and Vision with a Multimodal Skip-gram Model. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 153–163.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

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*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved Baselines with Visual Instruction Tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Chandler May, Alex Wang, Shikha Bordia, Samuel Bowman, and Rachel Rudinger. 2019. On Measuring Social Biases in Sentence Encoders. 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 622–628.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Dimitri Coelho Mollo and Raphaël Millière. 2023. The vector grounding problem. *arXiv preprint arXiv:2304.01481*.

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.

Friedemann Pulvermüller. 2018. Neural reuse of action perception circuits for language, concepts and communication. *Progress in neurobiology*, 160:1–44.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118.

Carina Silberer and Mirella Lapata. 2012. Grounded models of semantic representation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1423–1433. Association for Computational Linguistics.

Carina Silberer and Mirella Lapata. 2014. Learning grounded meaning representations with autoencoders. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 721–732. Association for Computational Linguistics.

Robert A Stanfield and Rolf A Zwaan. 2001. The effect of implied orientation derived from verbal context on picture recognition. *Psychological science*, 12(2):153–156.

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.

Hao Tan and Mohit Bansal. 2020. Vokenization: Improving Language Understanding with Contextualized, Visual-Grounded Supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2066–2080.

Marco Tettamanti, Giovanni Buccino, Maria Cristina Saccuman, Vittorio Gallese, Massimo Danna, Paola Scifo, Ferruccio Fazio, Giacomo Rizzolatti, Stefano F Cappa, and Daniela Perani. 2005. Listening to action-related sentences activates fronto-parietal motor circuits. *Journal of cognitive neuroscience*, 17(2):273–281.

Wessel O Van Dam, Margriet Van Dijk, Harold Bekkering, and Shirley-Ann Rueschemeyer. 2012. Flexibility in embodied lexical-semantic representations. *Human brain mapping*, 33(10):2322–2333.

Gabriella Vigliocco, Pamela Perniss, and David Vinson. 2014. Language as a multimodal phenomenon: implications for language learning, processing and evolution.

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*.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-Scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Miaoran Zhang, Marius Mosbach, David Adelani, Michael Hedderich, and Dietrich Klakow. 2022. MCSE: Multimodal Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5959–5969, Seattle, United States. Association for Computational Linguistics.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

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

Rolf A Zwaan and Lawrence J Taylor. 2006. Seeing, acting, understanding: Motor resonance in language comprehension. *Journal of Experimental Psychology: General*, 135(1):1.

Appendix

A Sentence Templates

The neutral sentence templates where the word stimuli were embedded in order to obtain contextualised representations from the computational models were the following:

```
Someone mentioned the <word>.
The post was about the <word>.
Everyone was talking about the <word>.
They were all interested in the <word>.
People know about the <word>.
```

In one of our additional experiments (see Section 5.1), we used caption-like sentences to check whether they were more in-distribution for vision-language models and, therefore, yielded more EXP48- and brain-aligned representations. Below, we report the caption-like templates used for each word sub-category.

Templates used for the sub-category *food*:

```
There is a <word> on a table in a restaurant.
A <word> is on a kitchen table.
A woman is eating a <word>.
A <word> with a few glasses around.
A close-up of a <word>.
```

Templates used for the sub-category *vehicle*:

```
There is one man in a <word>.
A <word> is surrounded by a few people.
A woman is posing next to a <word>.
A <word> with a young man next to it.
A close-up of a <word>.
```

Templates used for the sub-category *tool*:

```
There is a man holding a <word>.
A <word> is lying on the ground.
A woman is using a <word>.
A <word> with some people in the background.
A close-up of a <word>.
```

Templates used for the sub-category *animal*:

```
There is a <word> eating voraciously.
A man is feeding a <word>.
A woman next to a <word>.
A <word> with a little girl staring at it.
A close-up of a <word>.
```

Templates used for the sub-category *negative event*:

```
There is a crowd looking scared because of a <word>.
Many people are trying to shelter from a <word>.
A <word> happening in a big city.
A <word> with many people involved.
A picture of a <word>.
```

Templates used for the sub-category *social event*:

```
There is a small crowd attending a <word>.
A few people are gathered for a <word>.
A <word> attended by a large group of people.
A <word> with many people involved.
A picture of a <word>.
```

Templates used for the sub-category *communication*:

```
There is a small crowd at a <word>.
A few people are participating in a <word>.
A <word> in a crowded room.
A <word> with many people involved.
A picture of a <word>.
```

Templates used for the sub-category *sound*:

```
There is a man hearing a <word>.
A few people seem to hear a <word>.
A <word> is heard by a few people.
A <word> with a few people listening to it.
A picture of a <word>.
```

B Additional RSA Results

B.1 Single-word vs. contextualised representations

Our choice to derive word representations by including them in sentences was guided by the intuition that single words could have been an out-of-distribution input for computational models trained to output contextualised word representations. We empirically verified that representations obtained by embedding words within templates yield higher alignment than those obtained by passing single words to the models. We show the EXP48 and brain alignment obtained with both embedding-extraction procedures in Figure 6.

B.2 Layer-wise RSA results

In the main paper, we reported RSA results calculated from model representations averaged across the three layers yielding the highest alignment individually. Here, we provide a layer-wise visualisation of RSA results, which allows observing how EXP48 vs. brain alignment changes throughout model layers. Specifically, layer-wise Spearman correlations against EXP48 are displayed in Figure 7, while those against fMRI responses are in Figure 8.

B.3 RSA with additional baselines

For completeness, in Table 1 we report RSA results including three additional models: CLIP (Radford et al., 2021), a vision-language model pretrained contrastively on 400M image-caption pairs, and the distributional models GloVE (Pennington et al., 2014) and Word2vec (Mikolov et al., 2013). The distributional models were originally included in Fernandino et al. (2022); note that the brain correlations we report differ from the ones from Fernandino et al. (2022), as they computed an average across participant-wise brain correlations, while we averaged brain RDMs across participants *before* computing correlations.

Model	ρ EXP48	ρ *Brain*
SimCSE	0.52	0.22
MCSE	0.45	0.19
CLAP	0.03	0.00
BERT	0.53	0.23
VisualBERT	0.27	0.12
CLIP	0.41	0.14
GloVe	0.45	0.14
Word2vec	0.42	0.125

Table 1: Spearman correlations quantifying the alignment of models' representational spaces with EXP48 and brain responses.

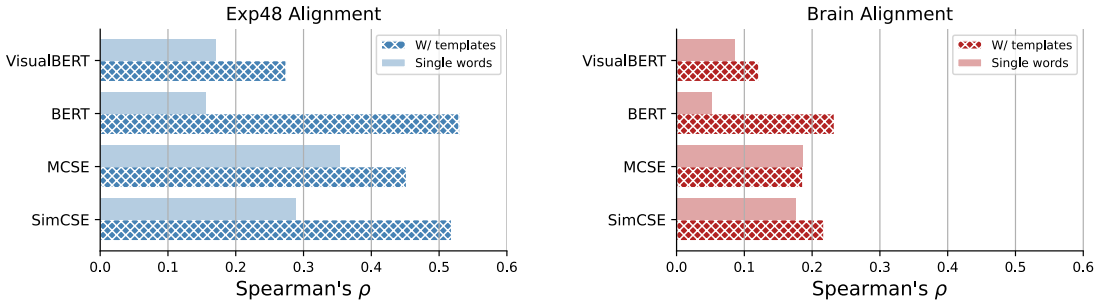

Figure 6: Spearman correlations observed from model representations obtained by passing single words vs. words embedded in templates. The left-hand panel shows the alignment with EXP48 and the right-hand one with brain responses.

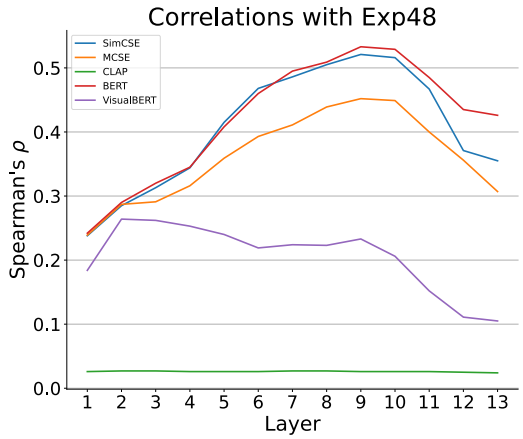

Figure 7: Spearman correlations indicating how representational similarity between model representations and EXP48 representations changes along model layers.

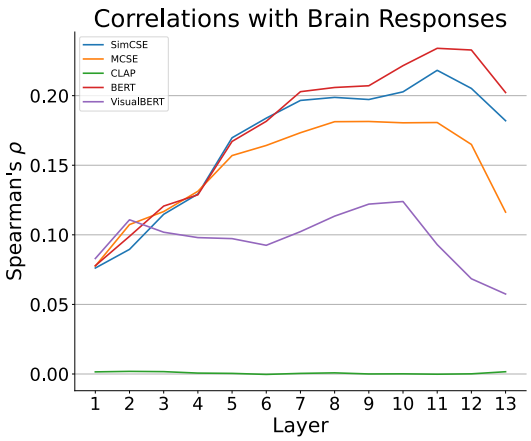

Figure 8: Spearman correlations indicating how representational similarity between model representations and brain responses changes along model layers.