

Appendix

A Limitations

Despite our efforts to develop a robust experimental design, a few limitations should still be considered when drawing conclusions from our results. First, the experiential model we used—while showed to reflect brain-relevant information—relies on pre-defined dimensions that are, to some extent, abstract and high-level. In this sense, it may fail to capture low-level perceptual information relevant for modelling human word representations and, perhaps, learnt by computational models.

Another limitation is that our study focuses on single words as opposed to longer text passages. To some extent, our experiments suggest that this may affect *machine* language processing, as we found representations by computational models to be more aligned with brain responses when words are passed within sentence templates vs. in isolation (see also Appendix C.1). Similarly, the isolated-word setup is likely to play a role *human* language processing—as insightfully observed by Zwaan (2014), context is crucial in determining the mental simulations people engage during language comprehension and the amount of perceptual detail they contain.

Lastly, two additional limitations are related to the models included in our study. The first is that, while SimCSE, MCSE and CLAP are comparable in their learning objectives and architecture, they still differ in the amount of fine-tuning data. While, in principle, this could be problematic, we observed that SimCSE—despite being fine-tuned on *less* sentences than MCSE—still proves to be *more* EXP48- and brain-aligned. As for CLAP, we acknowledge that the smaller amount of fine-tuning audio-caption pairs, together with the fact that it did not optimise a SimCSE-like objective jointly with the CLIP-like one, may have played a role in its poor performance.

The second limitation related to our model selection is that we only considered architectures fine-tuned contrastively, which are known to reflect high-level correspondences between objects present in images and captions, while failing to capture more fine-grained relations between modalities (Hendricks and Nematzadeh, 2021; Parcalabescu et al., 2022; Thrush et al., 2022; Liu et al., 2023; Chen et al., 2023; Bavaresco et al., 2024). Therefore, it could be that different trends would emerge if repeating the experiments with computational

models trained with different objectives.

B 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>.

C Additional RSA Results

C.1 Single-word vs. contextualised representations

Our choice to derive word representations by including them in sentences was guided by the intu-

<i>Model</i>	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.

ition 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 1.

C.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 2, while those against fMRI responses are in Figure 3.

C.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.

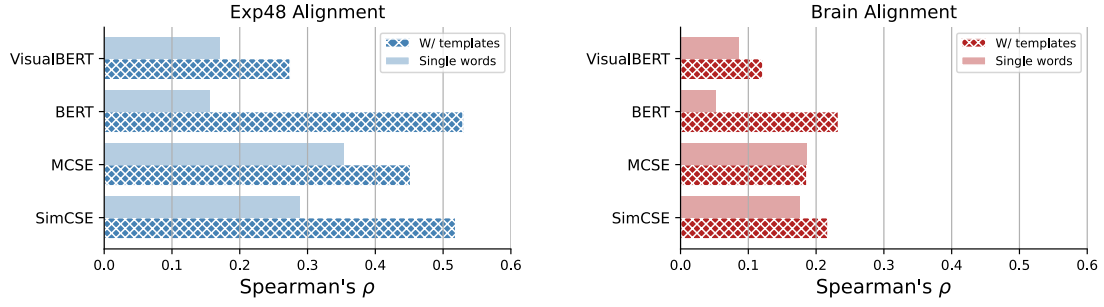


Figure 1: Spearman correlations observed from model representations obtained by passing single words vs. words embedded in templates. The left-hand panel shows the alignment with EXP48 and the right-hand one with brain responses.

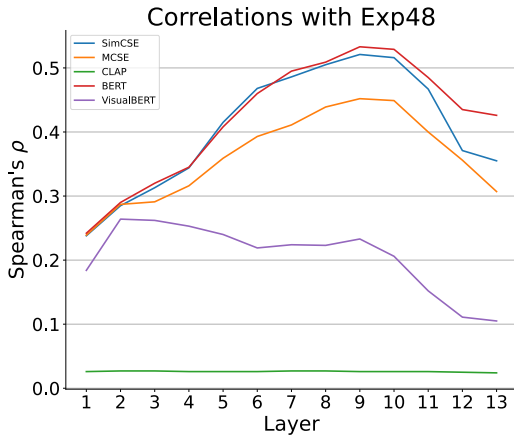


Figure 2: Spearman correlations indicating how representational similarity between model representations and EXP48 representations changes along model layers.

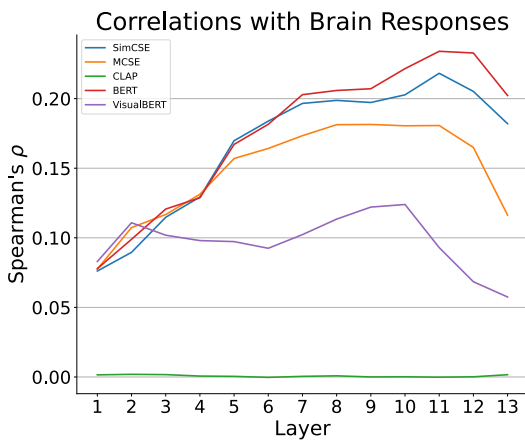


Figure 3: Spearman correlations indicating how representational similarity between model representations and brain responses changes along model layers.

References

- Anna Bavaresco, Alberto Testoni, and Raquel Fernández. 2024. [Don't Buy it! Reassessing the Ad Understanding Abilities of Contrastive Multimodal Models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 870–879, Bangkok, Thailand. Association for Computational Linguistics.
- Xinyi Chen, Raquel Fernández, and Sandro Pezzelle. 2023. [The BLA Benchmark: Investigating Basic Language Abilities of Pre-Trained Multimodal Models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5817–5830, Singapore. 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.
- 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.
- Fangyu Liu, Guy Emerson, and Nigel Collier. 2023. [Visual Spatial Reasoning](#). *Transactions of the Association for Computational Linguistics*, 11:635–651.
- 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*.
- 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.

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.

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.

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