MODALITY-AGNOSTIC FMRI DECODING OF VISION AND LANGUAGE

Mitja Nikolaus*
CerCo, CNRS
Toulouse, France
mitja.nikolaus@cnrs.fr

Milad Mozafari*
Torus AI
Toulouse, France
milad.mozafari@torus-actions.fr

Nicholas Asher
IRIT, Université Paul Sabatier
Toulouse, France
nicholas.asher@irit.fr

Leila Reddy & Rufin VanRullen
CerCo, CNRS
Toulouse, France
{leila.reddy,rufin.vanrullen}@cnrs.fr

ABSTRACT

Previous studies have shown that it is possible to map brain activation data of subjects viewing images onto the feature representation space of not only *vision* models (modality-specific decoding) but also *language* models (cross-modal decoding). In this work, we introduce and use a new large-scale fMRI dataset ($\sim 8,500$ trials per subject) of people watching both images and text descriptions of such images. This novel dataset enables the development of *modality-agnostic* decoders: a single decoder that can predict which stimulus a subject is seeing, irrespective of the modality (image or text) in which the stimulus is presented. We train and evaluate such decoders to map brain signals onto stimulus representations from a large range of publicly available vision, language and multimodal (vision+language) models. Our findings reveal that (1) modality-agnostic decoders perform as well as (and sometimes even better than) modality-specific decoders (2) modality-agnostic decoders mapping brain data onto representations from unimodal models perform as well as decoders relying on multimodal representations (3) while language and low-level visual (occipital) brain regions are best at decoding text and image stimuli, respectively, high-level visual (temporal) regions perform well on both stimulus types.

1 INTRODUCTION

Advances in deep-learning-based computational models of language and vision paired with large-scale open source fMRI datasets have fostered the development of brain decoding models which classify or reconstruct stimuli that subjects were seeing based on their brain activations (Naselaris et al., 2009; Nishimoto et al., 2011; Pereira et al., 2018; VanRullen & Reddy, 2019; Ozcelik & VanRullen, 2023; Scotti et al., 2023; Tang et al., 2023b; Benchetrit et al., 2023; Karamolegkou et al., 2023; Xia, 2024). A range of studies have presented *modality-specific* mappings between fMRI brain activation data of subjects viewing stimuli in one modality (e.g. images) and feature representation space of models of the same modality (e.g. vision models). More recently, it has been shown that these mappings can also be trained in a cross-modal fashion, i.e. mappings between fMRI data from one modality and feature space of models of another modality (e.g. between fMRI data of subjects viewing images and representations from language models) (Matsuo et al., 2017; Takada et al., 2020; Huang et al., 2021; Ferrante et al., 2023; Tang et al., 2023a).

Here, we present a new fMRI dataset and use it to develop *modality-agnostic* decoders. A modality-agnostic decoder is trained on fMRI data from multiple modalities (here: vision and language) and can retrieve the stimulus (image or caption) a subject is seeing irrespective of the modality. In contrast to *modality-specific* decoders that can be applied only in the single modality that they were

*Joint first authors

trained on, *modality-agnostic* decoders can be applied in multiple modalities, even without knowing the stimulus modality a priori.

The fMRI experiment consists of 6 subjects viewing $\sim 8,500$ stimuli (images and captions) while performing a one-back cross-modal matching task. An additional set of 70 images and 70 captions were presented to all subjects and serves as a test set for the decoders. This new fMRI dataset will be released publicly in an upcoming publication.

We train modality-agnostic decoders based on this new multimodal fMRI dataset and evaluate them on their decoding performance in both modalities. Our results show that modality-agnostic decoders generally perform on par with their respective modality-specific counterparts, despite the additional challenge of uncertainty about the stimulus modality. We further compare decoders trained on features extracted from a range of vision, language and multimodal models and show that multimodal representations do not increase decoding performance above that of decoders trained on unimodal representations in the correct modality. Finally, an ROI-based analysis reveals that activity from high-level visual brain regions is most effective for training modality-agnostic decoders, suggesting that these regions contain representations that are to some degree "amodal".

2 METHODS

2.1 fMRI EXPERIMENT

Six subjects (2 female, age between 20 and 50 years, all right-handed) participated in the experiment after providing informed consent. The study was performed in accordance with French national ethical regulations (Comité de Protection des Personnes, ID 2019-A01920-57). We collected functional MRI data using a 3T Philips ACHIEVA scanner. At the start of each session, we further acquired high-resolution anatomical images for each subject. Scanning was spanned over 10 sessions (except for sub-01: 11 sessions), each consisting of 13-16 runs during which the subjects were presented 86 stimuli. The stimulus type varied randomly between images and captions. Each stimulus was presented for 2.5 seconds at the center of the screen, the inter-stimulus interval was 1s. Further details on the scanner configuration and experimental setup are reported in Appendix A.1.

Subjects were performing a one-back matching task: They were instructed to press a button whenever the stimulus was matching the immediately preceding one. In case the previous stimulus was of the same modality (e.g. two captions in a row), the subjects were instructed to press a button if the stimuli were matching exactly. In the cross-modal case (e.g. an image followed by a caption), the button had to be pressed if the caption was a valid description of the image, and vice versa. Positive one-back trials occurred on average every 10 stimuli.

Images and captions were taken from the training and validation sets of the COCO dataset (Lin et al., 2014, COCO contains 5 matching captions for each image, of which we only considered the shortest one in order to fit on the screen and to ensure a comparable length for all captions). As our training set, a random subset of images and another random subset of captions was selected for each subject. All these stimuli were presented only a single time. Additionally, a shared subset of 140 stimuli (70 images and 70 captions) was presented repeatedly (on average: 26 times, min: 22, max: 31) to each subject in order to reduce noise, serving as our test set. These stimuli were inserted randomly between the training stimuli. Note that for each image (respectively, caption) presented to the subject, we retained the corresponding caption (resp. image) in order to estimate model features in the opposite modality (e.g. language model features for an image stimulus).

2.2 fMRI PREPROCESSING

Preprocessing of the fMRI data was performed using SPM 12 (Ashburner et al., 2014). We applied Slice Time Correction and Realignment for each subject. Each session was coregistered with the subject's T1 scan. Afterwards, we transformed all data to the MNI305 space (Evans et al., 1993) using Freesurfer (Fischl, 2012), and explicit gray matter masks were created using SPM and applied for each subject.

We fit a first GLM for each subject on data from all sessions. We included regressors for events that re-occurred across runs and sessions, i.e. test images, test captions, fixations, blank screens,

and one-back trials. The residual volumes of these GLMs were the inputs to a second-phase GLM, which was fit for each run separately, and intended to derive single-trial beta-values. We included regressors for each training image and caption presented during the run. As output of these second GLMs we obtained a single volume of beta-values for each training caption and image. One-back target trials were excluded from the second-phase GLM.

2.3 MODALITY-AGNOSTIC DECODERS

We trained regression models that take fMRI beta-values from training stimuli (images and captions) as input and predict latent representations extracted from vision, language, and multimodal models.

We consider as vision models: ResNet (He et al., 2016), ViT (Dosovitskiy et al., 2020), and DINOv2 (Oquab et al., 2023); as language models: BERT (Devlin et al., 2019), GPT2 (Radford et al., 2019), Llama2 (Touvron et al., 2023), mistral and mixtral (Jiang et al., 2023). Regarding multimodal models, we extract features from VisualBERT (Li et al., 2019), BridgeTower (Xu et al., 2023), LXMERT (Tan & Bansal, 2019), ViLT (Kim et al., 2021), CLIP (Radford et al., 2021), ImageBind (Girdhar et al., 2023), Flava (Singh et al., 2022). For each target stimulus (image or caption), we extracted model features from the corresponding image for vision models, the corresponding caption for language models, and a concatenated representation of both image and caption for the multimodal models. We use publicly available pretrained models implemented in the HuggingFace Transformers library (Wolf et al., 2020). In order to estimate the effect of model training, we further extract features from a randomly initialized Flava model as a baseline. Further details on feature extraction and decoder training can be found in Appendix A.2.

The decoders were linear ridge-regression models implemented using scikit-learn (Pedregosa et al., 2011). The regularization hyperparameter α was optimized using 5-fold cross validation on the training set (values considered: $\alpha \in \{1e3, 1e4, 1e5, 1e6, 1e7\}$). Afterwards, a final model was trained using the best α on the whole training set.

Finally, the models were evaluated on the held-out test data (140 stimuli, 70 captions and 70 images) using pairwise accuracy calculated using cosine distance. In the case of cross-modal decoding (e.g. mapping an image stimulus into the latent space of a language model), a trial was counted as correct if the caption corresponding to the image (according to the ground-truth in COCO) was closest.

3 RESULTS

3.1 MODALITY-AGNOSTIC DECODING

We present a comparison of pairwise accuracy scores for captions and images of modality-agnostic and modality-specific decoders in Figure 1. Results for individual subjects can be found in Appendix A.5. Generally, we observe that modality-agnostic decoders perform as well as the modality-specific decoders trained on the correct modality, and much better than the modality-specific decoders trained on the opposite modality. They achieve this high performance despite the additional challenge of not knowing the modality of the stimulus the subject was seeing.

When calculating the overall modality-agnostic decoding performance as average performance for captions and images (bottom panel of Figure 1), we find that modality-agnostic decoders based on the best multimodal features (ViLT: 0.88 ± 0.03) do not perform substantially better than decoders based on the best language features (GPT2-xl: 0.86 ± 0.03) and only slightly better than decoders trained on the best vision features (Dino-large: 0.83 ± 0.03).

3.2 ROI-BASED MODALITY-AGNOSTIC DECODING

The results presented so far are based on decoders trained on data from the whole brain. To provide insight into the organization of visual and language representations in the brain, we additionally trained decoders on subsets of voxels for 3 Regions Of Interest (ROI), defined based on an anatomical atlas (Destrieux et al., 2010): A low-level visual area spanning mainly the occipital lobe, a high-level visual area in the temporal lobe, and a left-lateralized language-related area broadly defined based on the findings of Fedorenko et al. (2010). Surface plots of these 3 ROIs are depicted in Figure 2. Further details on the ROI definition can be found in Appendix A.4.

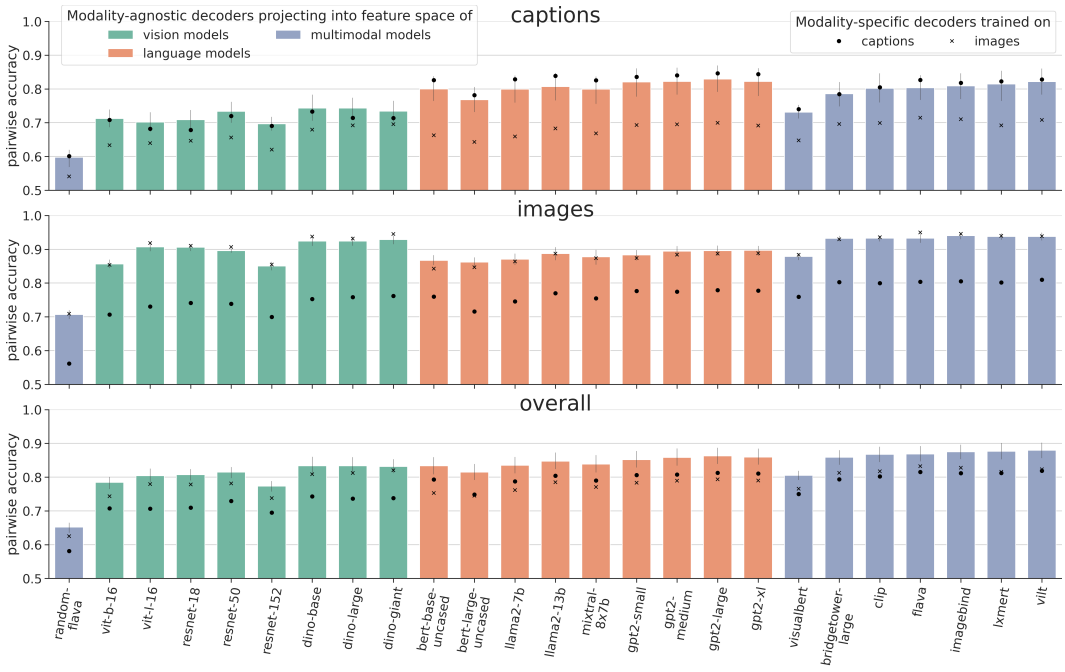

Figure 1: Decoding accuracy for captions (top), images (middle) and overall (bottom) for modality-agnostic decoders trained on full data (bars), compared to modality-specific decoders trained on either just linguistic fMRI data (•) or just on visual fMRI data (×). Error bars indicate 95% confidence intervals for modality-agnostic decoders. Chance performance is at 0.5.

Pairwise decoding accuracies for the ROI-based decoders are presented in Figure 3. Even though the ROI-based decoders rely on 20x less dimensions (∼10,000 voxels) than the whole brain decoders (∼215,000 voxels), image decoding performance of decoders based on the low-level visual ROI is on par with decoders that use the whole brain data, and caption decoding performance for the language ROI is close to it as well.

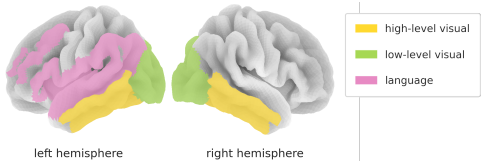

Figure 2: Surface plots of the 3 ROIs. The average numbers of voxels in the high-level visual area is 11,340; in the low-level visual area 10,578; and in the language area 11,193.

As expected, we find that both modality-agnostic and modality-specific decoders' decoding accuracy for captions is lowest in the low-level visual area and highest in the language area; for images, it is lowest in the language area and highest in the low-level visual area. However, decoders trained on high-level visual areas of the temporal cortex perform well, both for decoding images and captions, and are systematically the highest across both modalities (Fig 3, bottom). This suggests that representations in this area are to some degree amodal.

4 Discussion

In this study, we presented a novel large-scale fMRI dataset and used it to train modality-agnostic decoders for vision and language. The fMRI data is unique in that it contains a large number of *separate* trials for *matched* visual and language stimuli (images and captions from the COCO dataset). Previous studies that relied on unimodal fMRI data (e.g. Chang et al., 2019; Allen et al., 2022) required either manual annotations to map stimuli from multiple modalities into a shared semantic space (Popham et al., 2021) or training of linear transformation matrices based on additional multimodal paired training data (Tang et al., 2023a). Other multimodal fMRI datasets usually consist of *simultaneous* presentations of visual and language stimuli (e.g. movies Huth et al., 2012; Çukur et al., 2013; Cichy & Lahner, 2021), which allows for the study of multimodal feature integration

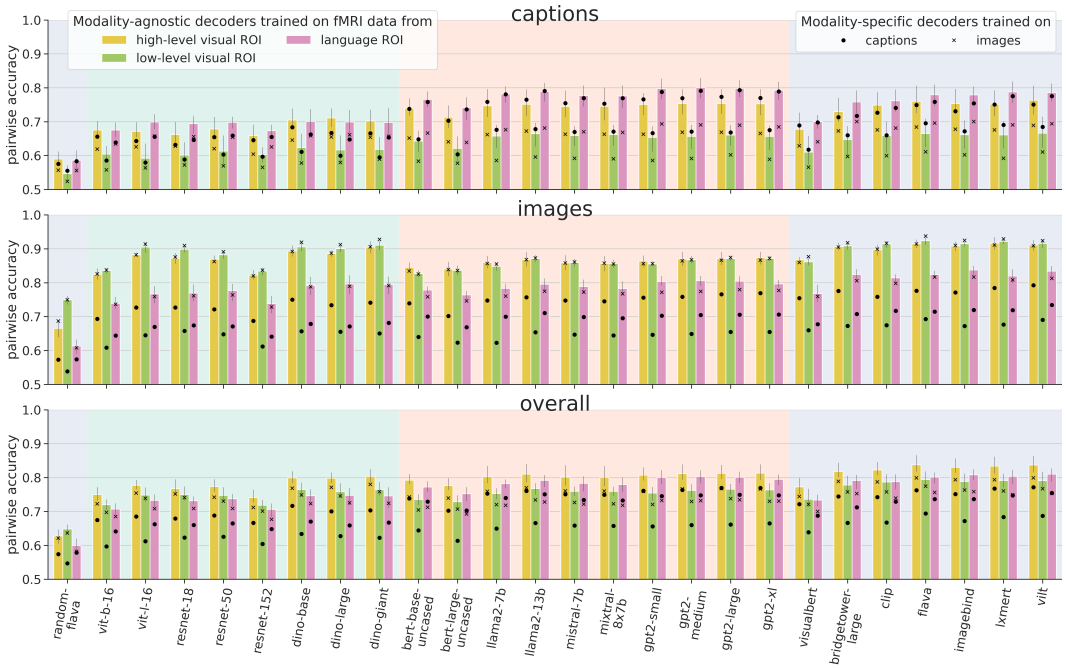

Figure 3: Decoding accuracy for captions (top), images (middle) and overall (bottom) for modality-agnostic (bars) and modality-specific decoders (• and ×) trained on 3 ROIs. The background colors reflect the model features that the decoders project into (vision, language or multimodal).

(Bonnici et al., 2016; Khosla et al., 2021; Dong & Toneva, 2023), but does not allow for the study of modalities in isolation.

The results of our decoding experiments based on this new dataset suggest that in order to build modality-agnostic decoders, we do not necessarily need representations from multimodal models; unimodal representations (especially from language models) can lead to comparably high performance. Two recent studies found that multimodal transformers (CLIP and BridgeTower) learn more aligned representations in language and vision than unimodal transformers (Wang et al., 2023; Tang et al., 2023a). In our study, we evaluated a large range of unimodal and multimodal representations, and found that especially representations extracted from more recent large language models (e.g. GPT2-xl) are as good as multimodal representations. Reasons for these different results could be that the aforementioned studies only considered language representations extracted from substantially smaller language models (BERT and RoBERTa (Liu et al., 2019)) and that models were compared in terms of their *encoding* performance, while we measured *decoding* performance (Kriegeskorte & Douglas, 2019).

Tang et al. (2023a) trained cross-modal encoding models between data from participants viewing movies and listening to audio books and found that "tuning for concepts in language and vision is positively correlated in most regions outside of visual cortex, it is negatively correlated in visual cortex." This phenomenon could explain why we do not observe higher performance of modality-agnostic decoders compared to modality-specific ones when trained on low-level visual ROIs: If the same stimuli presented in the visual and language modality are represented differently in these ROIs, training in one modality will not improve performance in the other modality.

Our ROIs contain several "amodal" regions that have been identified in previous studies (Devereux et al., 2013; Fairhall & Caramazza, 2013; Popham et al., 2021), such as the middle and inferior temporal gyrus (part of the high-level visual ROI) and the left angular gyrus and left posterior cingulate gyrus (language ROI). The superior performance of modality-agnostic decoders for these ROIs confirms that these regions share representations between modalities. In future work, we plan to perform a more fine-grained searchlight-based analysis to identify specific "amodal" regions, i.e. regions in which the performance advantage of modality-agnostic decoders is highest.

ACKNOWLEDGMENTS

This research was funded by grants from the French Agence Nationale de la Recherche (ANR: AI-REPS grant number ANR-18-CE37-0007-01 and ANITI grant number ANR-19-PI3A-0004) as well as the European Union (ERC Advanced grant GLoW, 101096017). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

REFERENCES

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

John Ashburner, Gareth Barnes, Chun-Chuan Chen, Jean Daunizeau, Guillaume Flandin, and Karl Friston. SPM12, 2014. URL https://www.fil.ion.ucl.ac.uk/spm/software/spm12/.

Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. Brain decoding: toward real-time reconstruction of visual perception, October 2023. URL http://arxiv.org/abs/2310.19812. arXiv:2310.19812 [cs, eess, q-bio].

Jason W. Bohland, Hemant Bokil, Cara B. Allen, and Partha P. Mitra. The Brain Atlas Concordance Problem: Quantitative Comparison of Anatomical Parcellations. *PLOS ONE*, 4(9):e7200, September 2009. ISSN 1932-6203. doi: 10.1371/journal.pone.0007200. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0007200. Publisher: Public Library of Science.

Heidi M. Bonnici, Franziska R. Richter, Yasemin Yazar, and Jon S. Simons. Multimodal Feature Integration in the Angular Gyrus during Episodic and Semantic Retrieval. *The Journal of Neuroscience*, 36(20):5462–5471, May 2016. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.4310-15.2016. URL https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.4310-15.2016.

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

Radoslaw Martin Cichy and Benjamin Lahner. The Algonauts Project 2021 Challenge, 2021.

Christophe Destrieux, Bruce Fischl, Anders Dale, and Eric Halgren. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage*, 53(1):1–15, October 2010. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2010.06.010. URL https://www.sciencedirect.com/science/article/pii/S1053811910008542.

Barry J. Devereux, Alex Clarke, Andreas Marouchos, and Lorraine K. Tyler. Representational Similarity Analysis Reveals Commonalities and Differences in the Semantic Processing of Words and Objects. *The Journal of Neuroscience*, 33(48):18906–18916, November 2013. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.3809-13.2013. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3852350/.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Dota Tianai Dong and Mariya Toneva. Vision-Language Integration in Multimodal Video Transformers (Partially) Aligns with the Brain, November 2023. URL http://arxiv.org/abs/2311.07766. arXiv:2311.07766 [cs].

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2020.

A.C. Evans, D.L. Collins, S.R. Mills, E.D. Brown, R.L. Kelly, and T.M. Peters. 3D statistical neuroanatomical models from 305 MRI volumes. In *1993 IEEE Conference Record Nuclear Science Symposium and Medical Imaging Conference*, pp. 1813–1817 vol.3, October 1993. doi: 10.1109/NSSMIC.1993.373602. URL https://ieeexplore.ieee.org/abstract/document/373602.

S. L. Fairhall and A. Caramazza. Brain Regions That Represent Amodal Conceptual Knowledge. *Journal of Neuroscience*, 33(25):10552–10558, June 2013. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.0051-13.2013. URL https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.0051-13.2013.

Evelina Fedorenko, Po-Jang Hsieh, Alfonso Nieto-Castañón, Susan Whitfield-Gabrieli, and Nancy Kanwisher. New Method for fMRI Investigations of Language: Defining ROIs Functionally in Individual Subjects. *Journal of Neurophysiology*, 104(2):1177–1194, August 2010. ISSN 0022-3077, 1522-1598. doi: 10.1152/jn.00032.2010. URL https://www.physiology.org/doi/10.1152/jn.00032.2010.

Matteo Ferrante, Furkan Ozcelik, Tommaso Boccato, Rufin VanRullen, and Nicola Toschi. Multimodal decoding of human brain activity into images and text. In *UniReps: the First Workshop on Unifying Representations in Neural Models*. arXiv, 2023. URL http://arxiv.org/abs/2305.11560. arXiv:2305.11560 [cs].

Bruce Fischl. FreeSurfer. *NeuroImage*, 62(2):774–781, 2012. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2012.01.021. Place: Netherlands Publisher: Elsevier Science.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One Embedding Space To Bind Them All. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15180–15190, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Girdhar_ImageBind_One_Embedding_Space_To_Bind_Them_All_CVPR_2023_paper.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Wei Huang, Hongmei Yan, Kaiwen Cheng, Chong Wang, Jiyi Li, Yuting Wang, Chen Li, Chaorong Li, Yunhan Li, Zhentao Zuo, and Huafu Chen. A neural decoding algorithm that generates language from visual activity evoked by natural images. *Neural Networks*, 144:90–100, December 2021. ISSN 0893-6080. doi: 10.1016/j.neunet.2021.08.006. URL https://www.sciencedirect.com/science/article/pii/S0893608021003117.

Alexander G. Huth, Shinji Nishimoto, An T. Vu, and Jack L. Gallant. A Continuous Semantic Space Describes the Representation of Thousands of Object and Action Categories across the Human Brain. *Neuron*, 76(6):1210–1224, December 2012. ISSN 08966273. doi: 10.1016/j.neuron.2012.10.014. URL https://linkinghub.elsevier.com/retrieve/pii/S0896627312009348.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B, October 2023. URL http://arxiv.org/abs/2310.06825. arXiv:2310.06825 [cs].

Antonia Karamolegkou, Mostafa Abdou, and Anders Søgaard. Mapping Brains with Language Models: A Survey. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 9748–9762, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.618. URL https://aclanthology.org/2023.findings-acl.618.

Meenakshi Khosla, Gia H. Ngo, Keith Jamison, Amy Kuceyeski, and Mert R. Sabuncu. Cortical response to naturalistic stimuli is largely predictable with deep neural networks. *Science Advances*, 7(22):eabe7547, May 2021. doi: 10.1126/sciadv.abe7547. URL https://www.science.org/doi/10.1126/sciadv.abe7547. Publisher: American Association for the Advancement of Science.

Wonjae Kim, Bokyung Son, and Ildoo Kim. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 5583–5594. PMLR, July 2021. URL https://proceedings.mlr.press/v139/kim21k.html. ISSN: 2640-3498.

Katarzyna Krasnowska-Kieraś and Alina Wróblewska. Empirical Linguistic Study of Sentence Embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5729–5739, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1573. URL https://www.aclweb.org/anthology/P19-1573.

Nikolaus Kriegeskorte and Pamela K Douglas. Interpreting encoding and decoding models. *Current Opinion in Neurobiology*, 55:167–179, April 2019. ISSN 0959-4388. doi: 10.1016/j.conb.2019.04.002. URL https://www.sciencedirect.com/science/article/pii/S0959438818301004.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. VisualBERT: A Simple and Performant Baseline for Vision and Language. *arXiv:1908.03557 [cs]*, August 2019. URL http://arxiv.org/abs/1908.03557. arXiv: 1908.03557.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, volume 8693, pp. 740–755. Springer International Publishing, Cham, 2014. doi: 10.1007/978-3-319-10602-1_48. URL http://link.springer.com/10.1007/978-3-319-10602-1_48. Series Title: Lecture Notes in Computer Science.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach, 2019. URL https://arxiv.org/abs/1907.11692.

Eri Matsuo, Ichiro Kobayashi, Shinji Nishimoto, Satoshi Nishida, and Hideki Asoh. Describing Semantic Representations of Brain Activity Evoked by Visual Stimuli. In *31st Conference on Neural Information Processing Systems*. arXiv, 2017. URL http://arxiv.org/abs/1802.02210. arXiv:1802.02210 [cs].

Thomas Naselaris, Ryan J. Prenger, Kendrick N. Kay, Michael Oliver, and Jack L. Gallant. Bayesian Reconstruction of Natural Images from Human Brain Activity. *Neuron*, 63(6):902–915, September 2009. ISSN 08966273. doi: 10.1016/j.neuron.2009.09.006. URL https://linkinghub.elsevier.com/retrieve/pii/S0896627309006850.

Shinji Nishimoto, An T. Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L. Gallant. Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies. *Current Biology*, 21(19):1641–1646, October 2011. ISSN 0960-9822. doi: 10.1016/j.cub.2011.08.031. URL https://www.cell.com/current-biology/abstract/S0960-9822(11)00937-7. Publisher: Elsevier.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut,

Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision, April 2023. URL http://arxiv.org/abs/2304.07193. arXiv:2304.07193 [cs].

Furkan Ozcelik and Rufin VanRullen. Natural scene reconstruction from fMRI signals using generative latent diffusion, June 2023. URL http://arxiv.org/abs/2303.05334. arXiv:2303.05334 [cs, q-bio].

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, and Vincent Dubourg. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

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

Sara F. Popham, Alexander G. Huth, Natalia Y. Bilenko, Fatma Deniz, James S. Gao, Anwar O. Nunez-Elizalde, and Jack L. Gallant. Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nature Neuroscience*, 24(11):1628–1636, November 2021. ISSN 1097-6256, 1546-1726. doi: 10.1038/s41593-021-00921-6. URL https://www.nature.com/articles/s41593-021-00921-6.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. URL https://insightcivic.s3.us-east-1.amazonaws.com/language-models.pdf.

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. Learning Transferable Visual Models From Natural Language Supervision. *Proceedings of the 38th International Conference on Machine Learning*, pp. 16, 2021.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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)*, pp. 3980–3990, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://www.aclweb.org/anthology/D19-1410.

Mehraveh Salehi, Abigail S. Greene, Amin Karbasi, Xilin Shen, Dustin Scheinost, and R. Todd Constable. There is no single functional atlas even for a single individual: Functional parcel definitions change with task. *NeuroImage*, 208:116366, March 2020. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2019.116366. URL https://www.sciencedirect.com/science/article/pii/S1053811919309577.

Paul S Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Ethan Cohen, Aidan J Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth A Norman, and Tanishq Mathew Abraham. Reconstructing the Mind's Eye: fMRI-to-Image with Contrastive Learning and Diffusion Priors. In *NeurIPS*, 2023.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. FLAVA: A Foundational Language and Vision Alignment Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15638–15650, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Singh_FLAVA_A_Foundational_Language_and_Vision_Alignment_Model_CVPR_2022_paper.html.

Saya Takada, Ren Togo, Takahiro Ogawa, and Miki Haseyama. Generation of Viewed Image Captions From Human Brain Activity Via Unsupervised Text Latent Space. In *2020 IEEE International Conference on Image Processing (ICIP)*, pp. 2521–2525, October 2020. doi: 10.1109/ICIP40778.2020.9191262. URL https://ieeexplore.ieee.org/abstract/document/9191262. ISSN: 2381-8549.

Hao Tan and Mohit Bansal. 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)*, pp. 5100–5111, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1514. URL https://aclanthology.org/D19-1514.

Jerry Tang, Meng Du, Vy A. Vo, Vasudev Lal, and Alexander G. Huth. Brain encoding models based on multimodal transformers can transfer across language and vision. In *Thirty-seventh Conference on Neural Information Processing Systems*. arXiv, 2023a. URL https://openreview.net/forum?id=UPefaFqjNQ. arXiv:2305.12248 [cs].

Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G. Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5):858–866, May 2023b. ISSN 1546-1726. doi: 10.1038/s41593-023-01304-9. URL https://www.nature.com/articles/s41593-023-01304-9. Number: 5 Publisher: Nature Publishing Group.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models, July 2023. URL http://arxiv.org/abs/2307.09288.

Rufin VanRullen and Leila Reddy. Reconstructing faces from fMRI patterns using deep generative neural networks. *Communications Biology*, 2(1):1–10, May 2019. ISSN 2399-3642. doi: 10.1038/s42003-019-0438-y. URL https://www.nature.com/articles/s42003-019-0438-y. Number: 1 Publisher: Nature Publishing Group.

Aria Y. Wang, Kendrick Kay, Thomas Naselaris, Michael J. Tarr, and Leila Wehbe. Natural language supervision with a large and diverse dataset builds better models of human high-level visual cortex, July 2023. URL https://www.biorxiv.org/content/10.1101/2022.09.27.508760v2. Pages: 2022.09.27.508760 Section: New Results.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, and Morgan Funtowicz. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pp. 38–45, 2020.

Weihao Xia. DREAM: Visual Decoding From Reversing Human Visual System. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

Xiao Xu, Chenfei Wu, Shachar Rosenman, Vasudev Lal, Wanxiang Che, and Nan Duan. BridgeTower: Building Bridges between Encoders in Vision-Language Representation Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9):10637–10647, June 2023. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v37i9.26263. URL https://ojs.aaai.org/index.php/AAAI/article/view/26263.

Tolga Çukur, Shinji Nishimoto, Alexander G. Huth, and Jack L. Gallant. Attention during natural vision warps semantic representation across the human brain. *Nature Neuroscience*, 16(6):763–770, June 2013. ISSN 1546-1726. doi: 10.1038/nn.3381. URL https://www.nature.com/articles/nn.3381. Number: 6 Publisher: Nature Publishing Group.

A Appendix

A.1 fMRI Experiment Details

The functional MRI data was collected using a 3T Philips ACHIEVA scanner (gradient echo pulse sequence, TR=2s, TE=10ms, 41 slices with a 32-channel head coil, slice thickness=3mm with 0.2mm gap, in-plane voxel dimensions 3×3mm). High-resolution anatomical images for each subject (1×1×1mm voxels, TR=8.13ms, TE=3.74ms, 170 sagittal slices) were acquired at the start of each session.

Each run started and ended with an 8s fixation period. The stimulus type varied randomly between images and captions. Each stimulus was presented for 2.5 seconds at the center of the screen (visual angle: 14.6 degrees), captions were displayed in white on a gray background (font: "Consolas"). The inter-stimulus interval was 1s. Every 10 stimuli there was a fixation trial that lasted for 2.5s. Every 5min there was a longer fixation trial for 16s.

Exact numbers of training stimuli presented for each subject can be found in Table 1.

Table 1: Number of training stimuli for each subject.

Subject	# Stimuli
sub-01	9856
sub-02	8232
sub-03	8008
sub-04	8680
sub-05	8568
sub-06	8568

A.2 Feature extraction details

Pretrained models were taken from Huggingface or from their respective authors' repositories. Model versions for unimodal models are as indicated in Figure 1. For multimodal models, the exact version for CLIP was `clip-vit-large-patch14`, for ViLT `vilt-b32-mlm`, for LXMERT `lxmert-base-uncased`, for VisualBERT `visualbert-nlvr2-coco-pre`, for Imagebind `imagebind_huge`, and for Flava `flava-full`.

We extracted language features from all models by averaging the outputs for each token, as this has established as common practice for the extraction of sentence embeddings from Transformer-based language models (e.g. Krasnowska-Kieraś & Wróblewska, 2019; Reimers & Gurevych, 2019).

For Transformer-based vision models, we compare representations extracted by averaging the outputs for each patch with representations extracted from `[CLS]` tokens in Figure 4. We find that for almost all models, the mean features allow for higher decoding accuracies. For all experiments reported in the main paper we therefore only considered this method.

For multimodal models, we concatenated the vision and language features to create the final multimodal feature representation. We also trained decoders on only the language or vision features of the multimodal models. Their performance was in most cases comparable or worse than for the concatenated features, therefore we do not report them in the main text. Results using these features can however be found in Appendix A.3.

The models Flava and BridgeTower also allow for a direct extraction of multimodal features, we found however that they perform much worse than concatenated vision and language features and therefore did not consider these further in our experiments.

A.3 Results for vision and language features of multimodal models

In the main results, we only consider concatenated vision and language features of the multimodal models. We can however also just use the vision or language features of these models.

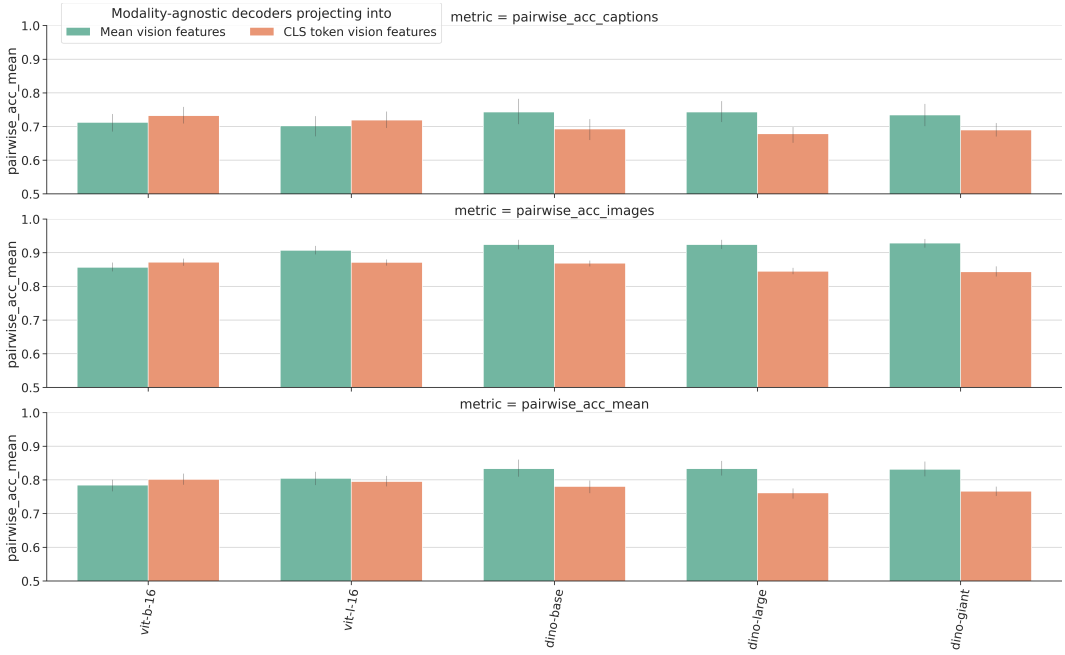

Figure 4: Pairwise accuracy for decoders based on vision features extracted by averaging the last hidden states ("Mean vision features") compared to when using features extracted from `[CLS]` tokens.

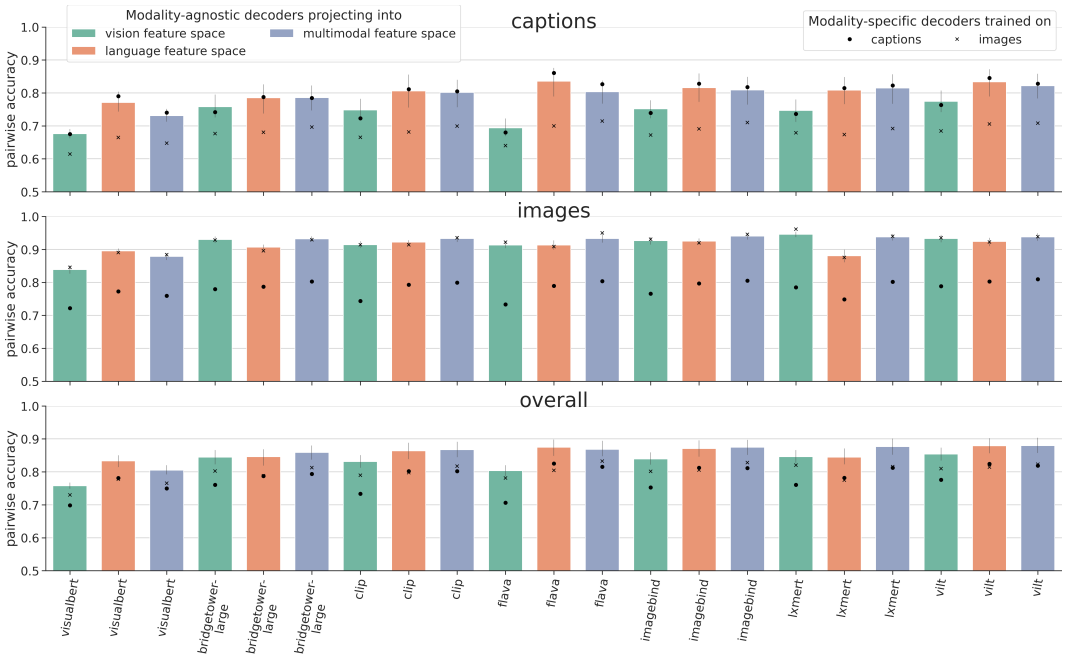

Figure 5: Pairwise accuracy for decoders based on vision, language, and multimodal features extracted from multimodal models.

Figure 5 compares the performance of decoders based on these different features for multimodal models. We find that the concatenated multimodal and language features usually perform best, in line with the main results comparing unimodal and multimodal features in Figure 1.

A.4 ROI DETAILS

We defined ROIs based on the anatomical Destrieux Atlas (Destrieux et al., 2010). The exact labels and names for each region included in each ROI can be found in Tables 2, 3, and 4. We defined non-overlapping regions of comparable size (in terms of number of voxels). While these ROI definitions allow us to perform a first analyses of differences in decoding performance in broadly-defined functional regions of the brain, this analysis suffers from the limitations that there is no universally agreed upon functional atlas of the human brain, the exact location of functional regions might also depend on the task, and there is substantial between-subject variability (Bohland et al., 2009; Salehi et al., 2020). In the future we plan to address these shortcomings by leveraging a more bottom-up approach in the form of searchlight analyses.

Table 2: Regions that were included in the high-level visual ROI

ID	Label	Names
21	L G_oc-temp_lat-fusifor	Lateral occipito-temporal gyrus (fusiform gyrus, O4-T4)
21	R G_oc-temp_lat-fusifor	Lateral occipito-temporal gyrus (fusiform gyrus, O4-T4)
23	L G_oc-temp_med-Parahip	Parahippocampal gyrus, parahippocampal part of the medial occipito-temporal gyrus, (T5)
23	R G_oc-temp_med-Parahip	Parahippocampal gyrus, parahippocampal part of the medial occipito-temporal gyrus, (T5)
61	L S_oc-temp_med_and_Lingual	Medial occipito-temporal sulcus (collateral sulcus) and lingual sulcus
61	R S_oc-temp_med_and_Lingual	Medial occipito-temporal sulcus (collateral sulcus) and lingual sulcus
60	L S_oc-temp_lat	Lateral occipito-temporal sulcus
60	R S_oc-temp_lat	Lateral occipito-temporal sulcus
37	L G_temporal_inf	Inferior temporal gyrus (T3)
38	L G_temporal_middle	Middle temporal gyrus (T2)
72	L S_temporal_inf	Inferior temporal sulcus
37	R G_temporal_inf	Inferior temporal gyrus (T3)
38	R G_temporal_middle	Middle temporal gyrus (T2)
72	R S_temporal_inf	Inferior temporal sulcus

Table 3: Regions that were included in the low-level visual ROI

ID	Label	Names
2	L G_and_S_occipital_inf	Inferior occipital gyrus (O3) and sulcus
19	L G_occipital_middle	Middle occipital gyrus (O2, lateral occipital gyrus)
20	L G_occipital_sup	Superior occipital gyrus (O1)
42	L Pole_occipital	Occipital pole
57	L S_oc_middle_and_Lunatus	Middle occipital sulcus and lunatus sulcus
58	L S_oc_sup_and_transversal	Superior occipital sulcus and transverse occipital sulcus
59	L S_occipital_ant	Anterior occipital sulcus and preoccipital notch (temporo-occipital incisure)
65	L S_parieto_occipital	Parieto-occipital sulcus (or fissure)
2	R G_and_S_occipital_inf	Inferior occipital gyrus (O3) and sulcus
19	R G_occipital_middle	Middle occipital gyrus (O2, lateral occipital gyrus)
20	R G_occipital_sup	Superior occipital gyrus (O1)
42	R Pole_occipital	Occipital pole
57	R S_oc_middle_and_Lunatus	Middle occipital sulcus and lunatus sulcus
58	R S_oc_sup_and_transversal	Superior occipital sulcus and transverse occipital sulcus
59	R S_occipital_ant	Anterior occipital sulcus and preoccipital notch (temporo-occipital incisure)
65	R S_parieto_occipital	Parieto-occipital sulcus (or fissure)
22	L G_oc-temp_med-Lingual	Lingual gyrus, ligual part of the medial occipito-temporal gyrus
22	R G_oc-temp_med-Lingual	Lingual gyrus, ligual part of the medial occipito-temporal gyrus

Table 4: Regions that were included in the language ROI

ID	Label	Names
12	L G_front_inf-Opercular	Opercular part of the inferior frontal gyrus
13	L G_front_inf-Orbital	Orbital part of the inferior frontal gyrus
14	L G_front_inf-Triangul	Triangular part of the inferior frontal gyrus
25	L G_pariet_inf-Angular	Angular gyrus
15	L G_front_middle	Middle frontal gyrus (F2)
34	L G_temp_sup-Lateral	Lateral aspect of the superior temporal gyrus
36	L G_temp_sup-Plan_tempo	Planum temporale or temporal plane of the superior temporal gyrus
35	L G_temp_sup-Plan_polar	Planum polare of the superior temporal gyrus
4	L G_and_S_subcentral	Subcentral gyrus (central operculum) and sulci
26	L G_pariet_inf-Supramar	Supramarginal gyrus
9	L G_cingul-Post-dorsal	Posterior-dorsal part of the cingulate gyrus (dPCC)
10	L G_cingul-Post-ventral	Posterior-ventral part of the cingulate gyrus (vPCC, isthmus of the cingulate gyrus)

A.5 PER-SUBJECT RESULTS

Results for individual subjects can be found in Figure 6. Among all subjects, we found similar converging results for decoding accuracies when comparing models, feature modalities, and modality-agnostic with modality-specific decoders.

Figure 6: Pairwise accuracy per subject