Modality-Agnostic Decoding of Vision and Language from fMRI

Mitja Nikolaus[1*†], **Milad Mozafari**[2†], **Isabelle Berry**[1], **Nicholas Asher**[3], **Leila Reddy**[1], **Rufin VanRullen**[1]

***For correspondence:** mitja.nikolaus@cnrs.fr (MN)

[†]These authors contributed equally to this work

[1]Université de Toulouse, CNRS, CerCo, Toulouse, France; [2]Torus AI, Toulouse, France; [3]Université de Toulouse, IRIT, Toulouse, France

Abstract Humans perform tasks involving the manipulation of inputs regardless of how these signals are perceived by the brain, thanks to representations that are invariant to the stimulus modality. In this paper, we present modality-agnostic decoders that leverage such modality-invariant representations to predict which stimulus a subject is seeing, irrespective of the modality in which the stimulus is presented. Training these modality-agnostic decoders is made possible thanks to our new large-scale fMRI dataset SemReps-8K, released publicly along with this paper. It comprises 6 subjects watching both images and short text descriptions of such images, as well as conditions during which the subjects were imagining visual scenes. We find that modality-agnostic decoders can perform as well as modality-specific decoders, and even outperform them when decoding captions and mental imagery. Further, a searchlight analysis revealed that large areas of the brain contain modality-invariant representations. Such areas are also particularly suitable for decoding visual scenes from the mental imagery condition.

Introduction

Several regions in the human brain have developed a high degree of specialization for particular lower-level perceptive as well as higher-level cognitive functions (***Kanwisher, 2010***). For many higher-level functions, it is crucial to be able to manipulate inputs regardless of the modality in which a stimulus was perceived by the brain. Such manipulations can be performed thanks to representations that are abstracted away from particularities of specific modalities, and are therefore *modality-invariant*. A range of theories have been developed to explain how and where in the human brain such invariant representations are created (***Damasio, 1989***; ***Binder et al., 2009***; ***Martin, 2016***; ***Barsalou, 2016***; ***Ralph et al., 2017***).

In this paper, we aim to develop modality-agnostic decoders that are specifically trained to leverage such modality-invariant representation patterns in the brain. In contrast to *modality-specific* decoders that are trained to be applied only in the modality that they were trained on, *modality-agnostic* decoders can be applied to decode brain patterns from multiple modalities, even without knowing a priori the modality the stimulus was presented in.

Regarding the terminology that will be used in the following, we use the terms modality-agnostic and modality-specific to refer to the way decoders are trained. The term modality-dependent describes patterns in brain activity that are completely unrelated for stimuli presented in varying modalities. In contrast, modality-invariant patterns do contain a significant degree of shared representational structure between modalities (***Man et al., 2012***). Evidence suggests that many regions in the brain contain mixtures of both types of information: A single region can show activation patterns with contributions of multiple modality-dependent and modality-invariant features (***Liuzzi et al., 2020***; ***Dirani and Pylkkänen, 2024***). Modality-agnostic decoders are explicitly trained to

leverage the shared information in modality-invariant patterns of the brain activity.

In order to build modality-agnostic decoders, large multimodal neuroimaging datasets with well-controlled stimuli across modalities are required. While large-scale datasets exist for vision (***Huth et al., 2012***; ***Chang et al., 2019***; ***Allen et al., 2022***), language (***Brennan and Hale, 2019***; ***Nastase et al., 2021***; ***Schoffelen et al., 2019***; ***Tang et al., 2023a***), and video (naturalistic movies) (***Aliko et al., 2020***; ***Visconti di Oleggio Castello et al., 2020***; ***Boyle et al., 2020***), none of these contain a controlled set of equivalent stimuli that are presented separately in both modalities. For instance, the different modalities in movies (vision and language) are complementary but do not always carry the same semantics. Further, they are not presented separately but simultaneously, impeding a study of the respective activity pattern caused by each modality in isolation.

The analyses of this paper are based on **SemReps-8K**, a new large-scale multimodal fMRI dataset of 6 subjects each viewing more than 8,000 stimuli which are presented separately in one of two modalities, as images of visual scenes or as descriptive captions of such images. In addition, the dataset also contains 3 imagery conditions for each subject, where they had to imagine an visual scene based on a caption description they had received before the start of the fMRI experiment.

We exploit this new data to develop modality-agnostic decoders that are trained on brain imaging data from multiple modalities, which we demonstrate here for the case of vision and language. We find that modality-agnostic decoders trained on this dataset perform on par with their respective modality-specific counterparts for decoding images, despite the additional challenge of uncertainty about the stimulus modality. For decoding captions, the modality-agnostic decoders even outperform the respective modality-specific decoders (because the former, but not the latter, can leverage the additional training data from the other image modality).

Additionally, we use this novel kind of decoders for a searchlight analysis to localize regions with modality-invariant representations in the brain. Previous studies that aimed to localize modality-invariant patterns were based on limited and rather simple stimulus sets and did not always agree on the exact location and extent of such regions (e.g. ***Vandenberghe et al., 1996***; ***Shinkareva et al., 2011***; ***Devereux et al., 2013***; ***Fairhall and Caramazza, 2013***; ***Jung et al., 2018***). We design a searchlight analysis based on a combination of modality-agnostic decoders and cross-decoding. The results reveal that modality-invariant patterns can be found in a widespread left-lateralized network across the brain, encompassing virtually all regions that have been proposed previously.

Finally, we find that modality-agnostic decoders trained only on data with perceptual input also generalize to conditions during which the subjects were performing mental imagery, and outperform both kinds of modality-specific decoders on this task.

Related Work

Modality-invariant representations

Decades of neuro-anatomical research (e.g. based on clinical lesions) and electrophysiology in non-human and human primates, as well as modern experiments leveraging recent brain imaging techniques have provided evidence that the activation patterns in certain brain areas are modality-dependent; for example, the occipital cortex responds predominantly to visual stimulation (***Felleman and Van Essen, 1991***; ***Sereno et al., 1995***; ***Grill-Spector and Malach, 2004***), and a commonly left-lateralized network responds to language processing tasks (***Zola-Morgan, 1995***; ***Fedorenko et al., 2010***, ***2011***; ***Friederici, 2017***; ***Brennan, 2022***).

More recent research has started to focus on higher-level regions that respond with *modality-invariant* patterns, i.e., patterns that are abstracted away from any modality-dependent information. A modality-invariant region responds with similar patterns to input stimuli of the same meaning, even if they are presented in different modalities (e.g. the word “cat” and picture of a cat).

Several theories and frameworks on how the brain transforms inputs from different modalities into modality-invariant representations have been proposed.[1] The **convergence zones view** pro-

[1] Here, we do not attempt to disambiguate whether information is modality-invariant or completely abstracted away from

poses that modality-dependent information coming from sensory cortices is integrated in multiple convergence zones that are distributed across the cortex, predominantly in temporal and parietal lobes (***Damasio, 1989***; ***Tranel et al., 1997***; ***Meyer and Damasio, 2009***). These convergence zones are organized hierarchically, learned associations are used to create abstractions from lower-level to higher-level feature representations (***Simmons and Barsalou, 2003***; ***Meyer and Damasio, 2009***). A perceived stimulus first causes modality-dependent activity in the related low-level region (e.g. visual cortex), subsequently higher-level convergence zones serve as relays that cause associated activity in other regions of the brain (e.g. the language network) (***Meyer and Damasio, 2009***; ***Kiefer and Pulvermüller, 2012***). According to ***Binder*** (***2016***), the most high-level convergence zones can become so abstract that they are representing amodal symbols.

The **GRAPES framework** (Grounding Representations in Action, Perception, and Emotion Systems) also suggests that representations are distributed across temporal and parietal areas of the cortex. More specifically, they are hypothesized to be situated in areas connected to the perception and manipulation of the environment, as well as in the language system (***Martin, 2009***, ***2016***). According to this theory, conceptual knowledge is organized in domains: For example, semantic information related to object form and object motion is represented within specific visual processing systems, regardless of the stimulus modality, and both for perception as well as imagination.

The **hub-and-spoke theory** states that cross-modal interactions are mediated by a single modality-invariant hub, located in the anterior temporal lobes (***Rogers et al., 2004***; ***Lambon Ralph et al., 2006***; ***Patterson and Lambon Ralph, 2016***; ***Ralph et al., 2017***). The hub contains a "continuous distributed representation space that expresses conceptual similarities among items even though its dimensions are not independently interpretable" (***Frisby et al., 2023***, p. 262). The spokes form the links between the hubs and the modality-dependent association cortices. Most importantly, semantic representations are not solely based in the hub, for a given concept all spokes that are linked to modalities in which the concept can be experienced do contribute to the semantic representation. This explains why selective damage to spokes can cause category-specific deficits (***Pobric et al., 2010***).

This is conceptually similar to some aspects of the **Global Workspace Theory**, which assumes both a multimodal convergence of inputs towards a specific (network of) region(s), and the possibility of flexibly recruiting unimodal regions into this Global Workspace (***Baars, 1993***, ***2005***).

While these and other theories partly disagree on *how* modality-invariant information is represented in the brain, they agree that such information is distributed across the cortex, and possibly overlapping with the semantic network (***Binder et al., 2009***; ***Huth et al., 2012***; ***Andrews et al., 2014***; ***Zwaan, 2016***).

Decoding of vision and language from fMRI

Early approaches of brain decoding focused on identifying and reconstructing limited sets of simple visual stimuli (***Haxby et al., 2001***; ***Cox and Savoy, 2003***; ***Kay et al., 2008***; ***Naselaris et al., 2009***; ***Nishimoto et al., 2011***). Soon after, attempts to decode linguistic stimuli could identify single words and short paragraphs with the help of models trained to predict features extracted from word embeddings (***Pereira et al., 2018***).

More recently, large-scale open source fMRI datasets for both vision and language have become available (***Chang et al., 2019***; ***Allen et al., 2022***; ***Schoffelen et al., 2019***; ***Tang et al., 2023b***) and allowed for the training of decoding models for a larger range and more complex naturalistic stimuli with the help of features extracted from deep learning models. For example, modality-specific decoders for vision can be trained by mapping the brain activity of subjects viewing naturalistic images to feature representation spaces of computational models of the same modality (i.e. vi-

input modalities, described in the literature as modality-independent (***Dirani and Pylkkänen, 2024***), supramodal (***Sanchez et al., 2020***), transmodal (***Ralph et al., 2017***), amodal (***Fairhall and Caramazza, 2013***), or abstract/conceptual (***Binder, 2016***) (see also distinctions made in ***Barsalou, 2016***). For our goal of building modality-agnostic decoders, it is only important that information be represented in a way invariant to the stimulus modality. We will come back to this point in the Discussion section.

sion models) (***Shen et al., 2019***; ***Beliy et al., 2019***; ***Lin et al., 2022***; ***Takagi and Nishimoto, 2023***; ***Ozcelik and VanRullen, 2023***). Moreover, a range of studies provided evidence that certain representations can transfer between vision and language by evaluating decoders in a modality that they were not trained on (***Shinkareva et al., 2011***; ***Man et al., 2012***; ***Fairhall and Caramazza, 2013***; ***Simanova et al., 2014***; ***Jung et al., 2018***). The performance in such *cross-modal decoding* evaluations always lags behind when compared to within-modality decoding. One explanation is that modality-specific decoders are not explicitly encouraged to pick up on modality-invariant features during training, and modality-dependent features do not transfer to other modalities.[2]

To address this limitation, we here propose to directly train *modality-agnostic decoders*, i.e. models that are exposed to multiple stimulus modalities during training in order to make it more likely that they are leveraging representations that are modality-invariant. Training this kind of decoder is enabled by the multimodal nature of our fMRI dataset: The stimuli are taken from COCO, a multimodal dataset of images with associated descriptive captions (***Lin et al., 2014***). During the experiment the subjects are exposed to stimuli in both modalities (images and captions) in separate trials. The modality-agnostic decoders learn to map brain activity associated with a specific stimulus independent of its presentation modality (image or caption) to high-level features extracted from a deep learning model. After training, a single modality-agnostic decoder can be used to decode stimuli from multiple modalities, leveraging representations that are invariant to the stimulus modality.

Decoding of mental imagery

Apart from decoding perceived stimuli, it is also possible to decode representations when subjects were performing mental imagery, without being exposed to any perceptual input.

Different theories on mental imagery processes emphasize either the role of the early visual areas (***Kosslyn et al., 1999***; ***Pearson, 2019***) or the role of the high-level visual areas in the ventral temporal cortex and frontoparietal networks (***Spagna et al., 2021***; ***Hajhajate et al., 2022***; ***Liu et al., 2025***). There is evidence for both kinds of theories in the form of neuroimaging studies that used decoding to identify stimuli during mental imagery. Some of these found relevant patterns in the early visual cortex (***Albers et al., 2013***; ***Naselaris et al., 2015***), others highlighted the role of higher-level areas in the ventral visual processing stream (***Stokes et al., 2009***; ***Reddy et al., 2010***; ***Lee et al., 2012***; ***VanRullen and Reddy, 2019***; ***Boccia et al., 2019***) as well as the precuneus and the intraparietal sulcus (***Johnson and Johnson, 2014***). These discrepancies can possibly be explained by differences in experimental design: For example, the early visual cortex might only become involved if the task requires the imagination of high-resolution details, which are represented in lower levels of the visual processing hierarchy (***Kosslyn and Thompson, 2003***).

Crucially, it has been shown that decoders trained exclusively on trials *with* perceptual input can generalize to imagery trials (***Stokes et al., 2009***; ***Reddy et al., 2010***; ***Lee et al., 2012***; ***Johnson and Johnson, 2014***; ***Naselaris et al., 2015***), providing evidence that representations formed during perception overlap to some degree with representations formed during mental imagery (***Dijkstra et al., 2019***).

In our study, we explored to what extent these findings hold true for more varied and complex stimuli. Following previous approaches, we use decoders trained exclusively on trials where subjects were viewing images and captions, and evaluated them on their ability to decode the imagery trials. We additionally hypothesized that mental imagery should be at least as conceptual as it is sensory and therefore primarily recruit modality-invariant representations. Consequently, modality-agnostic decoders should be ideally suited to decode mental imagery and outperform modality-specific decoders on that task.

[2]An exception is the case of modality-specific decoders that are trained using features from the other modality. Such decoders might also rely on modality-invariant features, however, these might be less easily decodable than the modality-dependent features.

Localizing modality-invariant regions

The first evidence for the existence of modality-invariant regions came from observations of patients with lesions in particular cortical regions which lead to deficits in the retrieval and use of knowledge across modalities (***Warrington and Shallice, 1984***; ***Warrington and Mccarthy, 1987***; ***Gainotti, 2000***; ***Damasio et al., 2004***). Semantic impairments across modalities have also been observed in patients with the neurodegenerative disorder semantic dementia (***Warrington, 1975***; ***Snowden et al., 1989***; ***Jefferies et al., 2009***).

In early work exploring the possible locations of modality-invariant regions in healthy subjects, brain activity was recorded using imaging techniques while they were presented with a range of concepts in two modalities (e.g. words and pictures). Regions that were active during semantic processing of stimuli in the first modality were compared to regions that were active during semantic processing of stimuli in the second modality. The conjunction of these regions was proposed to be modality-invariant (***Vandenberghe et al., 1996***; ***Moore and Price, 1999***; ***Bright et al., 2004***).

While this methodology allows for the identification of candidate regions in which semantic processing of multiple modalities occurs, it can not be used to probe the information represented in these regions. In order to compare the information content (i.e. multivariate patterns) of brain regions, researchers have developed Representational Similarity Analysis (RSA, ***Kriegeskorte et al., 2008***) as well as encoding and decoding analyses (***Naselaris et al., 2011***). More specifically, RSA has been used to find modality-invariant regions by comparing activation patterns of a candidate region when subjects are viewing stimuli from different modalities (***Devereux et al., 2013***; ***Handjaras et al., 2016***; ***Liuzzi et al., 2017***). This comparison is performed in an indirect way, by measuring the correlation of dissimilarity matrices of activation patterns. In turn, cross-decoding analysis can be leveraged to identify modality-invariant regions by training a classifier to predict the category of a stimulus in a given modality, and then evaluating its performance to predict the category of stimuli that were presented in another modality (***Shinkareva et al., 2011***; ***Man et al., 2012***; ***Fairhall and Caramazza, 2013***; ***Simanova et al., 2014***; ***Jung et al., 2018***). However, all these studies relied on a predefined set of stimulus categories, and can therefore not be easily extended to more realistic and complex stimuli, as we perceive them in our everyday life.

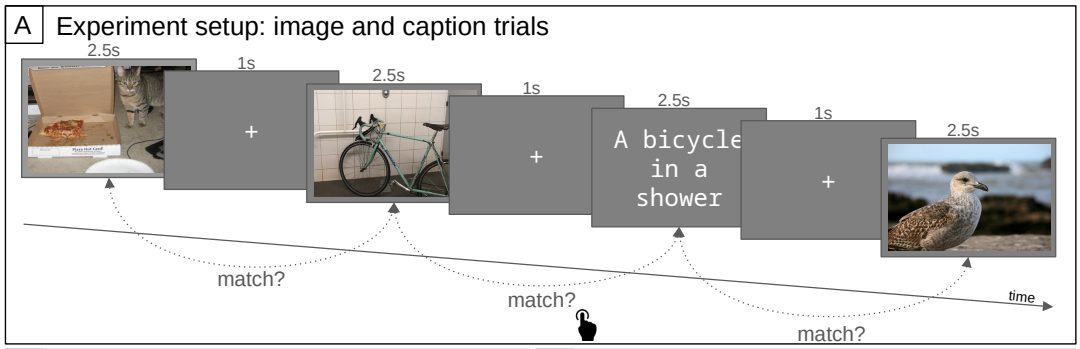
We summarize candidates for modality-invariant regions that have been identified by previous studies in Appendix 4. This overview reveals substantial disagreement regarding the possible locations of modality-invariant patterns in the brain. For example, ***Fairhall and Caramazza*** (***2013***) found modality-invariant representations in the left ventral temporal cortex (fusiform, parahippocampal, and perirhinal cortex), middle and inferior temporal gyrus, angular gyrus, parts of the prefrontal cortex as well as the precuneus. ***Shinkareva et al.*** (***2011***) found a larger network of left-lateralized regions, including additionally the left superior temporal, inferior parietal, supramarginal, inferior and inferior occipital, precentral and postcentral gyrus, supplementary motor area, intraparietal sulcus, cuneus, posterior cingulum as well as the right fusiform gyrus and the superior parietal gyrus, paracentral lobule on both hemispheres. In contrast, ***Jung et al.*** (***2018***) found modality-invariant representations only in the right prefrontal cortex. These diverging results can probably be explained by the limited number of stimuli as well as the use of artificially constructed stimuli in certain studies.

Recent advances in machine learning have enabled another generation of fMRI analyses based on large-scale naturalistic datasets. Here, we present a new multimodal dataset of subjects viewing both images and text. Most importantly, the dataset contains a large number of naturalistic stimuli in the form of complex visual scenes and full sentence descriptions of the same type of complex scenes, instead of pictures of single objects and words as commonly used in previous studies. This data enables the development of modality-agnostic decoders that are explicitly trained to leverage features that are shared across modalities. Further, we use this data to localize modality-invariant regions in the brain by applying decoders in a multimodal searchlight analysis.

Methods

fMRI Experiment

Six subjects (3 female, age between 22 and 47 years, all right-handed and fluent English speakers) 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 (gradient echo pulse sequence, TR=2s, TE=30ms, 46 slices with a 32-channel head coil, slice thickness=3mm with 0.2mm gap, in-plane voxel dimensions 3×3mm). At the start of each session, we further acquired high-resolution anatomical images for each subject (voxel size=1mm^3, TR=8.13ms, TE=3.74ms, 170 sagittal slices).



B Number of training stimuli

	Number of stimuli			
Subject	Images	Captions	Total	Overlapping
1	4897	4889	9786	211
2	4104	4108	8212	169
3	3992	3983	7975	153
4	4324	4329	8653	171
5	4266	4272	8538	181
6	4269	4267	8536	181

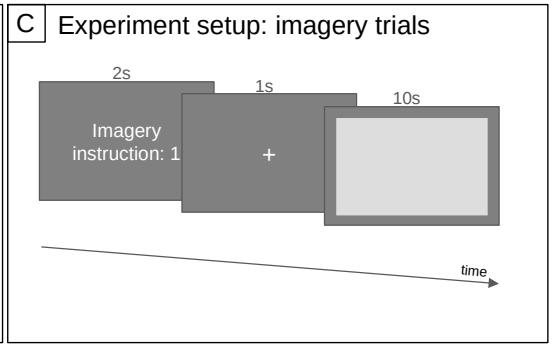


Figure 1. Panel A: Setup of the main fMRI experiment. Subjects were seeing images and captions in random alternation. Whenever the current stimulus matched the previous stimulus, the subjects were instructed to press a button (one-back matching task). Images and captions for illustration only; actual size and stimuli as described in the text. **Panel B:** Number of distinct training stimuli (excluding trials that were one-back targets or during which the subject pressed the response button). The number of overlapping stimuli indicates how many stimuli were presented both as caption and as image. There was an additional set of 140 stimuli (70 images and 70 captions) used for testing. **Panel C:** Setup of the fMRI experiment for the imagery trials. Subjects were instructed to remember 3 image descriptions with corresponding indices (numbers 1 to 3). One of these indices was displayed during the instruction phase, followed by a fixation phase, and then the subjects were imagining the visual scene for 10s.

Scanning was spanned over 10 sessions (except for sub-01: 11 sessions), each consisting of 13 to 16 runs during which the subjects were presented with 86 stimuli. Each run started and ended with an 8s fixation period. The stimulus type varied randomly inside each run 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 dark 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.

Subjects performed a one-back matching task: They were instructed to press a button whenever the stimulus matched the immediately preceding one (cf. Figure 1 Panel A). 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 matched 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***). This dataset 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. Spelling errors were corrected manually. As our training set, a random subset of images and another random subset of captions were selected for each subject. All these stimuli were presented only a single time.[3] Information on the number of training stimuli for each subject is shown in Figure 1 Panel B. Additionally, a shared subset of 140 stimuli (70 images and 70 captions) was presented repeatedly to each subject in order to reduce noise, serving as our test set (on average: 26 times, min: 22, max: 31).[4] These stimuli were inserted randomly between the training stimuli.

Note that for each stimulus presented to the subject (e.g. an image), we also have access to the corresponding stimulus in the other modality (the corresponding caption from the COCO dataset), allowing us to estimate model features based on both modalities (vision model features extracted from the image and language model features extracted from the corresponding caption) as well as multimodal features extracted from both the image and the caption.

In addition to these perceptual trials, there were 3 imagery trials for each subject (see also Figure 1 Panel C). Prior to the first fMRI scanning session, each subject was presented with a set of 20 captions (manually selected to be diverse and easy to visualize) that were not part of the perceptual trials, and they selected 3 captions for which they felt comfortable imagining a corresponding image. Then, they learned a mapping of each caption to a number (1, 2, and 3) so that they could be instructed to perform mental imagery of a specific stimulus, without having to present them with the caption again. The imagery trials occurred every second run, either at the beginning or the end of the run, so that each of the 3 imagery conditions were repeated on average 26 times (min: 23, max: 29). At the start of the imagery trial, the imagery instruction number was presented for 2s, then there was a 1s fixation period followed by the actual imagery period during which a light gray box was depicted for 10s on a dark gray background (the same background that was also used for trials with perceptual input). The light gray box was meant to represent the area in which the mental image should be "projected". At the end of the experiment, the subjects drew sketches of the images they had been imagining during the imagery trials.

We report dataset quality metrics on head motion and intersession alignment in Appendix 1.

fMRI Preprocessing

Preprocessing of the fMRI data was performed using SPM12 (***Ashburner et al., 2014***) via nipype (***Gorgolewski et al., 2011***). We applied slice time correction and realignment for each subject. Each session was coregistered with an anatomical scan of the respective subject's first session (downsampled to $2mm^3$). We created and applied explicit gray matter masks for each subject based on their anatomical scans using a maximally lenient threshold (probability>0).

In order to obtain beta-values for each stimulus, for each subject we fit a GLM (using SPM12) on data from all sessions. We included regressors for train images, train captions, test images, test captions, imagery trials, fixations, blank screens, button presses, and one-back target trials. One-back target trials as well as trials in which the participant pressed the button were excluded in the calculation of all training and test stimulus betas. As output of these GLMs we obtained beta-values for each training and test caption and image as well as the imagery trials.

Finally, we transformed the volume-space data to surface space Freesurfer (***Fischl, 2012***). We used trilinear interpolation and the fsaverage template in the highest possible resolution (163,842

[3] Except for the case of one-back target trials, for which the second presentation was however excluded during the preprocessing of the data.

[4] Contrary to the training stimuli which were randomly selected from the COCO dataset, the 70 test stimuli were chosen by hand to avoid including multiple scenes that could match the same semantic description. The 70 chosen images as well as their 70 corresponding captions constituted the test set.

vertices on each hemisphere) as target.

Modality-Agnostic Decoders

The multimodal nature of our dataset allowed for the training of modality-agnostic decoders. We trained decoders by fitting ridge regression models that take fMRI beta-values as input and predict latent representations extracted from a pretrained deep learning model. Further details on decoder training can be found in Figure 2 as well as Appendix 2.

While modality-specific decoders are trained only on brain imaging data of a single modality, modality-agnostic decoders are trained on brain imaging data from multiple modalities and therefore allow for decoding of stimuli irrespective of their modality.

More specifically, in our case the modality-specific decoders are trained on fMRI beta-values from one stimulus modality, e.g. when subjects were watching images (cf. Figure 2 panel A). Conversely, modality-agnostic decoders are trained jointly using fMRI data from both stimulus modalities (images and captions; cf. Figure 2 panel B). For all decoders, the features that serve as regression targets can either be unimodal (e.g. extracted from images using a vision model) or multimodal (e.g. extracted from both stimulus modalities using a multimodal model, cf. Figure 2 panel C).

We considered features extracted from a range of vision, language, and multimodal models: For vision features, we considered ResNet (***He et al., 2016***), ViT (***Dosovitskiy et al., 2020***), and DINOv2 (***Oquab et al., 2023***); for language features BERT (***Devlin et al., 2019***), GPT2 (***Radford et al., 2019***), Llama2 (***Touvron et al., 2023***), mistral and mixtral (***Jiang et al., 2023***). Regarding multimodal features, we extracted features from VisualBERT (***Li et al., 2019***), BridgeTower (***Xu et al., 2023***), ViLT (***Kim et al., 2021***), CLIP (***Radford et al., 2021***), ImageBind (***Girdhar et al., 2023***), Flava (***Singh et al., 2022***), Blip2 (***Li et al., 2023***), SigLip (***Zhai et al., 2023***), and Paligemma2 (***Steiner et al., 2024***). In order to estimate the effect of model training, we further extracted features from a randomly initialized ImageBind model as a baseline. Further details on feature extraction can be found in Appendix 2.

All decoders were evaluated on the held-out test data (140 stimuli, 70 captions and 70 images) using pairwise accuracy calculated using cosine distance. Prior to calculating the pairwise accuracy, the model predictions for all stimuli were standardized to have mean of 0 and standard deviation of 1.[5] 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.

Figure 3 provides an overview on the evaluation metrics. A modality-specific decoder for images can be evaluated on its ability to decode images (Panel A, top) and in a cross-decoding setup for captions (Panel B, bottom). In the same way, we can compute the respective evaluation metrics for modality-specific decoders trained on captions. For the case of modality-agnostic decoders, we evaluate performance for decoding both images and captions using the same single decoder that is trained on both modalities (Panel C).

Results

Modality-Agnostic Decoders

We first compared the performance of modality-specific and modality-agnostic decoders that are trained on the whole brain fMRI data based on different unimodal and multimodal features. The average pairwise accuracy scores are presented in Figure 4. Figure 5 presents pairwise accuracy scores separately for decoding images and for decoding captions. Results for individual subjects can be found in Appendix 6.

When analyzing the average decoding accuracy (Figure 4), we find that modality-agnostic decoders perform better than modality-specific decoders. To support this observation, we performed

[5] In the case of imagery decoding, the model predictions were standardized separately.

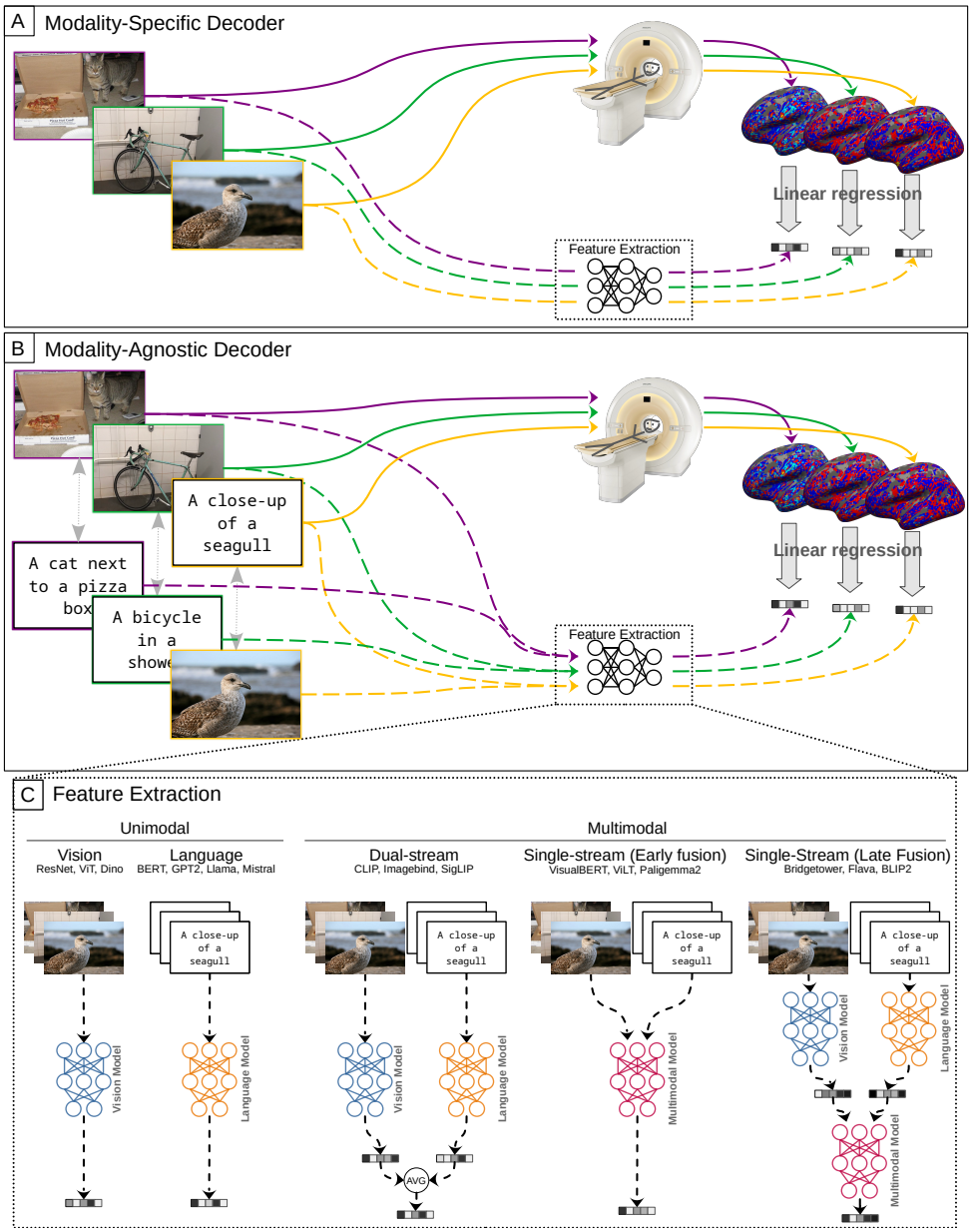


Figure 2. Training of modality-specific and modality-agnostic decoders. **Panel A:** Modality-specific decoders are trained on fMRI data of one modality (e.g. subjects viewing images) by mapping it to features extracted from the same stimuli. **Panel B:** Modality-agnostic decoders are trained jointly on fMRI data of both modalities (subjects viewing images and captions). **Panel C:** To train decoders, features can be either extracted unimodally from the corresponding images or captions, or by creating multimodal features based on both modalities. For example, to train a modality-agnostic decoder based on features from a unimodal language model, we map the fMRI data of subjects viewing captions to features extracted from the respective captions using this language model, as well as the fMRI data of subjects viewing images to features extracted by the language model from the corresponding captions. We can also train modality-specific decoders on features from another modality, for example by mapping fMRI data of subjects viewing images to features extracted from the corresponding captions using a language model (cf. crosses on orange bars in Figure 4 or using multimodal features (cf. crosses on blue bars in Figure 4).

two repeated measures ANOVAs (grouping the data by subject), once comparing the average accuracy values of modality-agnostic decoders (mean value: 81.86%) with those of modality-specific decoders trained on images (mean value: 78.15%), and once comparing modality-agnostic decoders

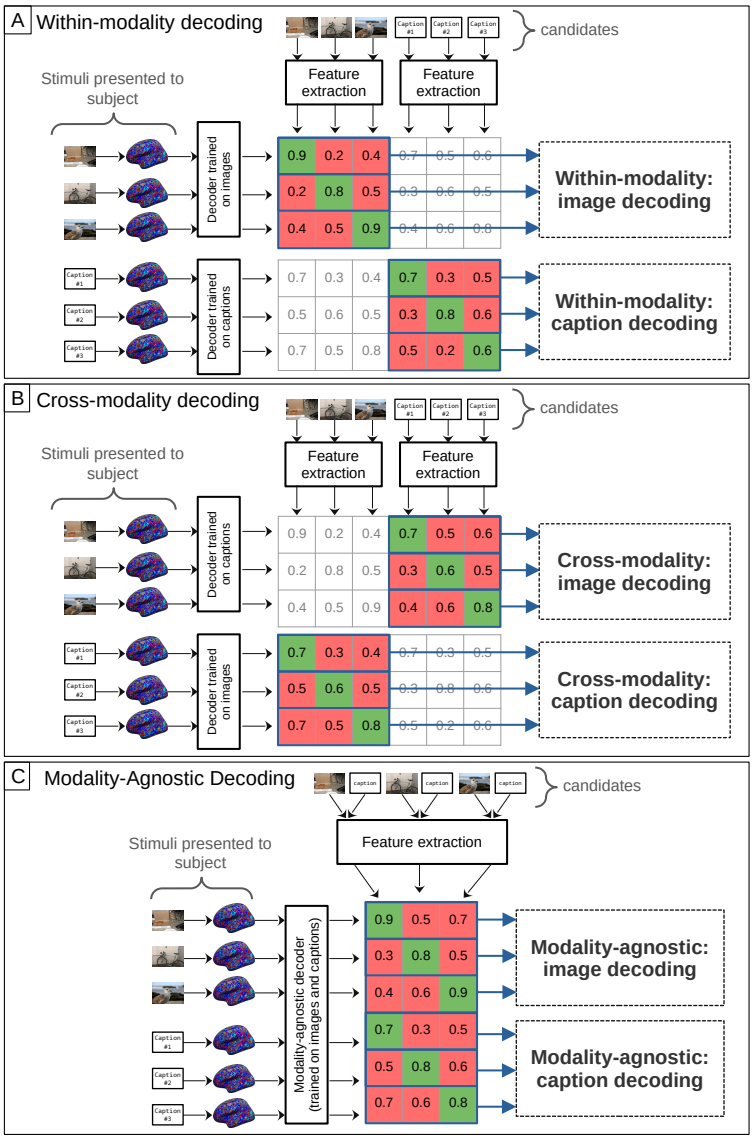


Figure 3. Evaluation of modality-specific and modality-agnostic decoders. The matrices display cosine similarity scores between features extracted from the candidate stimuli and features predicted by the decoder. The evaluation metric is pairwise accuracy, which is calculated row-wise: For a given matrix row, we compare the similarity score of the target stimulus on the diagonal (in green) with the similarity scores of all other candidate stimuli (in red). **Panel A:** Within-modality decoding metrics of modality-specific decoders. To compute within-modality accuracy for image decoding, a modality-specific decoder trained on images is evaluated on all stimuli that were presented as images. To compute within-modality accuracy for caption decoding, a modality-specific decoder trained on captions is evaluated on all caption stimuli. **Panel B:** Cross-modality decoding metrics of modality-specific decoders. To compute cross-modality accuracy for image decoding, a modality-specific decoder trained on captions is evaluated on all stimuli that were presented as images. To compute cross-modality accuracy for caption decoding, a modality-specific decoder trained on images is evaluated on all caption stimuli. **Panel C:** Metrics for modality-agnostic decoders. To compute modality-agnostic accuracy for image decoding, a modality-agnostic decoder is evaluated on all stimuli that were presented as images. The same decoder is evaluated on caption stimuli to compute modality-agnostic accuracy for caption decoding. Here we show feature extraction based on unimodal features for modality-specific decoders and based on multimodal features for the modality-agnostic decoder, in practice the feature extraction can be unimodal or multimodal for any decoder type (see also Figure 2).

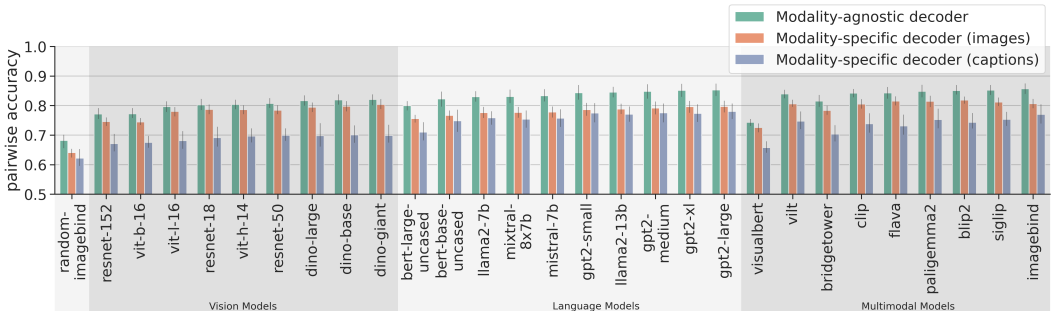


Figure 4. Average decoding scores for modality-agnostic decoders (green), compared to modality-specific decoders trained on data from subjects viewing images (orange) or on data from subjects viewing captions (purple). The metric is pairwise accuracy (see also Figure 3). Error bars indicate 95% confidence intervals calculated using bootstrapping. Chance performance is at 0.5.

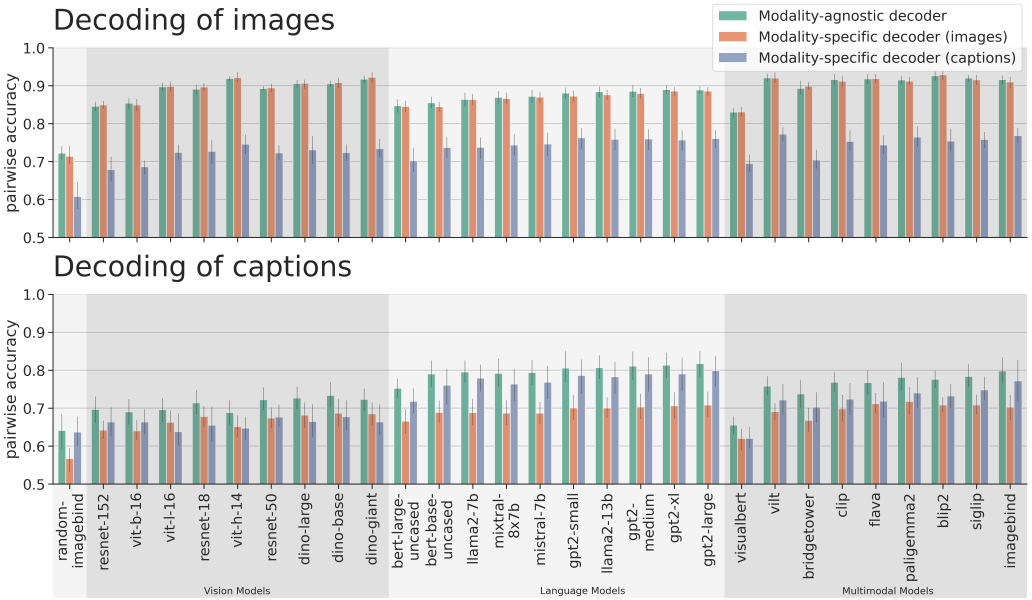


Figure 5. Decoding accuracy for decoding images (top) and for decoding captions (bottom). The orange bars in the top row indicate within-modality decoding scores for images, the purple bars in the bottom row indicate within-modality decoding scores for captions. The purple bars in the top row indicate cross-decoding scores for images, the orange bars in the bottom row indicate cross-decoding scores for captions (see also Figure 3). Error bars indicate 95% confidence intervals calculated using bootstrapping. Chance performance is at 0.5.

to modality-specific decoders trained on captions (mean value: 72.52%). In both cases, the accuracy values for the two decoder types were significantly different. When comparing modality-agnostic decoders to modality-specific decoders trained on images: decoder_type : $\beta = 0.037, SE = 0.004, p < 1 \cdot 10^{-18}$; and when comparing to modality-specific decoders trained on captions: decoder_type : $\beta = 0.093, SE = 0.005, p < 1 \cdot 10^{-93}$. This high performance (which can be attributed to the large training dataset used by modality-agnostic decoders) is achieved despite the additional challenge of not knowing the modality of the stimulus the subject was seeing.

Further, we observed that modality-agnostic decoders based on the best multimodal features (imagebind: $85.71\% \pm 2.58\%$) do not perform substantially better than decoders based on the best language features (GPT2-large: $85.31\% \pm 2.83\%$) and only slightly better than decoders trained on the best vision features (Dino-giant: $82.02\% \pm 2.43\%$). This result suggests that high-performing

modality-agnostic decoders do not necessarily need to rely on multimodal features; features extracted from language models can lead to equally high performance. When comparing the different architecture types of models for multimodal feature extraction (dual stream vs. single stream with early fusion vs. single stream with late fusion; cf. Panel C in Figure 2), we only observed a slight performance disadvantage for single-stream models with early fusion (Mean accuracy values for dual stream models: 85.04%; for single stream models with early fusion: 81.01%; for single stream models with late fusion 83.63%). We confirmed this observation with a repeated measures ANOVA (grouping the data by subject), comparing the decoding accuracy values of modality-agnostic decoders based on different families of multimodal features. The only significant effect was: model_family_single_stream_early_fusion : $\beta = -0.04, SE = 0.011, p < 1e-3$.

When analyzing the performance specifically for decoding images (Figure 5, top), we find that modality-agnostic decoders perform as well as the modality-specific decoders trained on images (green bars in the top row are at the same level as the orange bars in Figure 5) We found no statistically significant difference in their performances. A repeated measures ANOVA (grouping the data by subject), comparing the image decoding accuracy values of modality-agnostic decoders with those of modality-specific decoders trained on images resulted in a p-value of $p = 0.73$ for the effect of the decoder_type.

Further, modality-agnostic decoders even outperform modality-specific decoders trained on captions for decoding captions (purple bars in the bottom row are lower than green bars in Figure 5). This was supported by a repeated measures ANOVA (grouping the data by subject), comparing the caption decoding accuracy values of modality-agnostic decoders with those of modality-specific decoders trained on captions. The result was: decoder_type : $\beta = 0.036, SE = 0.006, p < 1 \cdot 10^{-8}$. In other words, even if we know that a brain pattern was recorded in response to the subject reading a caption, we are more likely to decode it accurately if we choose to apply a decoder trained using both modalities, than if we apply the appropriate decoder, trained only on captions.

Finally, we found that the cross-modal decoding performance for decoding visual stimuli (images) using decoders trained on captions (mean value: 73.31%) is higher than the cross-modal decoding performance in the other direction (mean value: 68.03%), corroborating similar results from ***Tang et al.*** (***2023a***) on movies and audio books (purple bars in top row are higher than orange bars in bottom row in Figure 5). We performed a repeated measures ANOVA (grouping the data by subject) to compare the pairwise accuracy values of both cross-decoding directions and found that the difference was significant: decoder_type : $\beta = 0.053, SE = 0.004, p < 1 \cdot 10^{-31}$ (the decoder_type variable is indicating whether the decoder is trained on images or captions).

Qualitative Decoding Results

To obtain a better understanding of the decoding performance of the modality-agnostic decoders, we inspected the decoding results for 5 randomly selected test stimuli. We created a large candidate set of 41,118 stimuli by combining the test stimuli and the training stimuli from all subjects. For each stimulus, we ranked these candidate set stimuli based on their similarity to the predicted feature vector. As the test stimuli were shared among all subjects, we could average the prediction feature vectors across subjects to obtain the best decoding results.

Figure 6 presents the results for decoding images using a modality-agnostic decoder trained on ImageBind features. We display the target stimulus along with the top-5 ranked candidate stimuli. We can observe some clear success cases (the train in the first row, a person eating pizza in the last row) but also failure cases (the teddy bear decoded as pizza). For the other stimuli, some aspects such as the high-level semantic class (e.g. vehicle, animal, sports) are correctly decoded: For the cars on the highway (5th row), the top-ranked images depict trains, which are also vehicles. For the dog (2nd row), the top images contain cats of similar colors. The footballer in the 6th row is decoded as a tennis player, the horses of the 7th row are decoded as zebras. Regarding the giraffe (3rd row), the model appears to have picked up on the fact that there was a body of water depicted in the image.

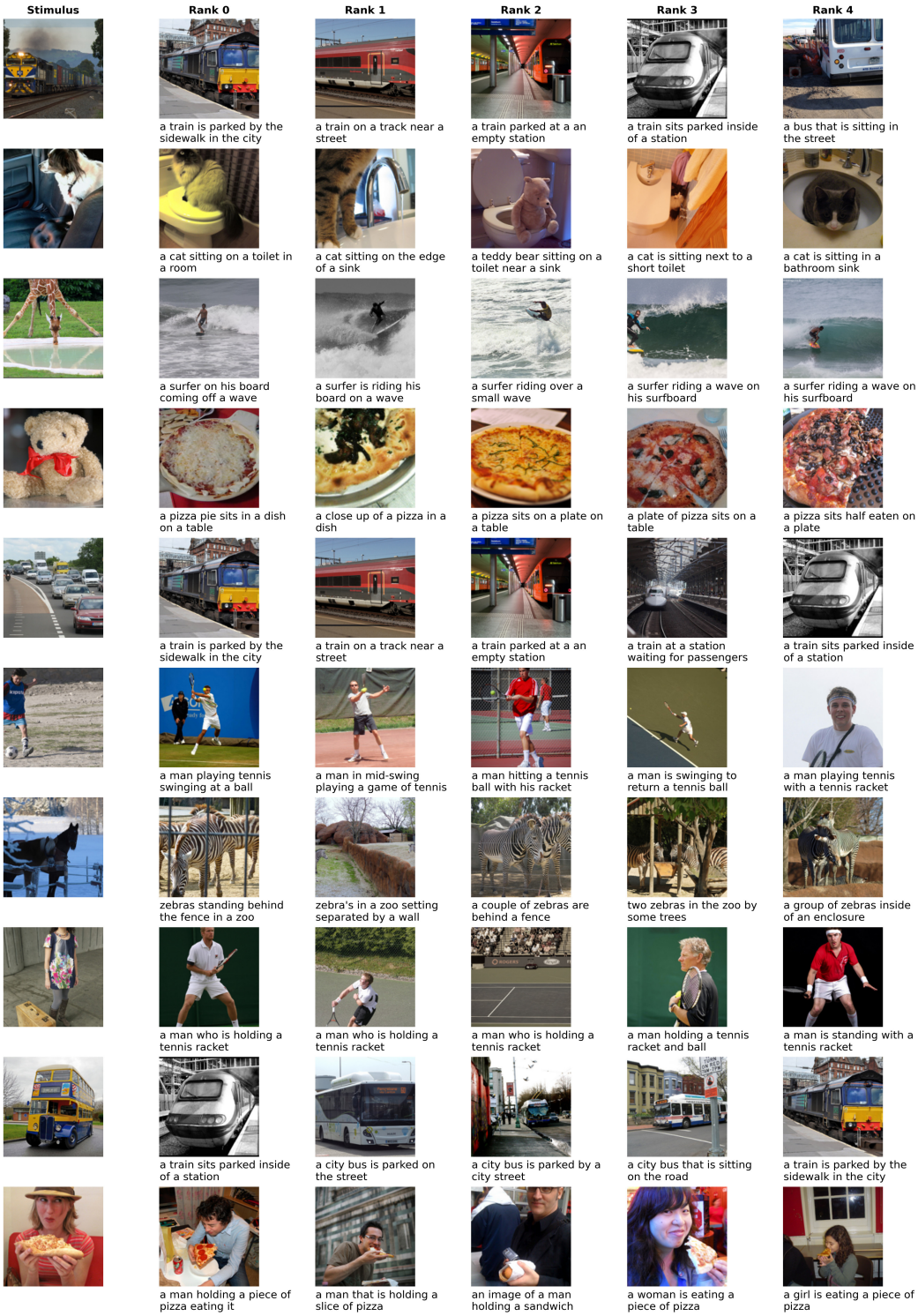


Figure 6. Decoding examples for image decoding using a modality-agnostic decoder. The first column shows the image the subject was seeing and the 5 following columns show the candidate stimuli with highest similarity to the predicted features, in descending order. We display both the image and the caption of the candidate stimuli because the decoder is based on multimodal features that are extracted from both modalities. All images were taken from the CoCo dataset (***Lin et al., 2014***).

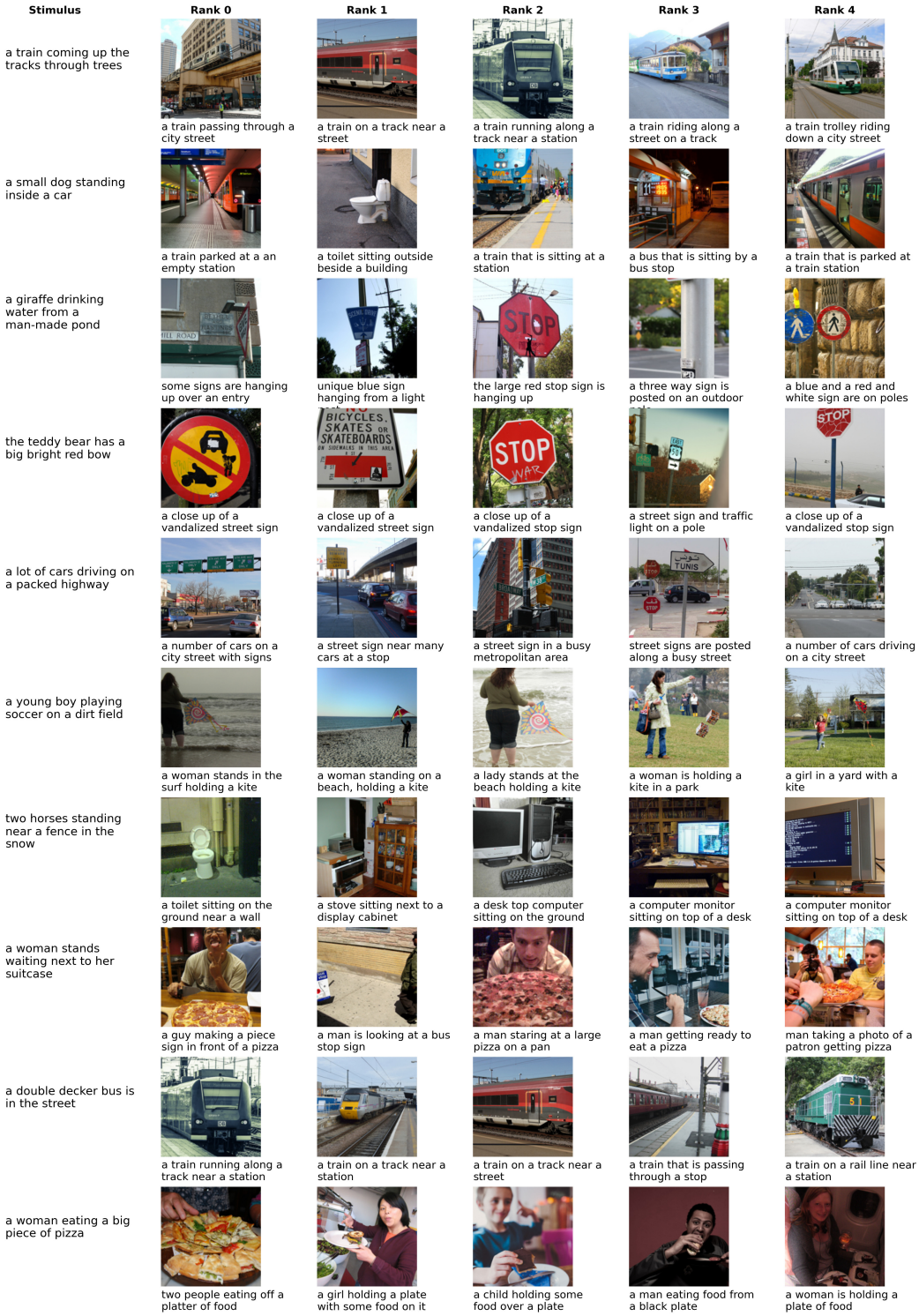


Figure 7. Decoding examples for caption decoding using a modality-agnostic decoder. For details see caption of Figure 6. All images were taken from the CoCo dataset (***Lin et al., 2014***).

Note that these qualitative results are not directly comparable with previous work on retrieval or reconstruction using the NSD dataset (***Allen et al., 2022***; ***Lin et al., 2022***; ***Takagi and Nishimoto, 2023***; ***Ozcelik and VanRullen, 2023***), as our data was collected on a 3T MRI scanner with lower

signal-to-noise-ratio than NSD's 7T MRI scanner.

The ranking results for decoding captions are depicted in Figure 7. The results are somewhat similar to the image decoding results, stimuli that were decoded successfully when presented as image such as the train are also decoded successfully when presented as caption; cases that were failures in the case of image decoding (e.g. the teddy bear) also fail here. However, the top-ranked stimuli for the dog (2nd row) do not always contain animals (the decoder seems to have picked up on the presence of a vehicle in the caption), and the horses of the 7th row do not get decoded correctly. The cars on the highway (5th row) get decoded more accurately than during image decoding.

We additionally provide qualitative results for modality-specific decoders in Appendix 3. These results generally reflect the observations from the quantitative results: Modality-agnostic decoders perform similarly to modality-specific decoders evaluated in a within-modality decoding setup, but substantially better than modality-specific decoders when evaluated in cross-decoding setups.

Modality-Invariant Regions

Modality-agnostic decoders perform best when they can leverage modality-invariant representations. To provide insight into the spatial organization of such modality-invariant representations in the brain, we performed a surface-based searchlight analysis.

Modality-invariant regions should contain patterns that generalize between stimulus modalities. Therefore, such regions should allow for decoding of stimuli in both modalities using a decoder that is trained to pick up on modality-invariant features, i.e. the decoding performance for images and captions of a modality-agnostic decoder should both be above chance. However, as a modality-agnostic decoder is trained on stimuli from both modalities, it could have learned to leverage certain features to project stimuli from one modality and different features to project stimuli from the other modality. We added two conditions to control that the representations directly transfer between the modalities by additionally training two modality-specific decoders and evaluating them according to their cross-decoding performance, i.e. we require that their decoding performance in the modality they were not trained on is above chance. These four conditions are summarized at the top of Figure 8.

We used ImageBind features for these searchlight analyses as they led to the highest decoding performance when using the whole brain data. The decoders were trained based on the surface projection of the fMRI beta-values. For each vertex, we defined a searchlight with a fixed size by selecting the 750 closest vertices, corresponding to an average radius of ~ 9.4mm.[6]

We trained and evaluated a modality-agnostic decoder and modality-specific decoders for both modalities on the beta-values for each searchlight location and each subject, providing us with a decoding accuracy scores for each location on the cortex. Then we performed one-tailed t-tests to identify locations in which the decoding performance is above chance ($acc > 0.5$). We aggregated all 4 comparisons by taking the minimum of the 4 t-values at each spatial location. Finally, we performed threshold-free cluster enhancement (TFCE, ***Smith and Nichols, 2009***) to identify modality-invariant ROIs (Figure 8, bottom).[7]

To estimate the statistical significance of the resulting clusters we performed a permutation test, combined with a bootstrapping procedure to estimate a group-level null distribution (see also ***Stelzer et al., 2013***). For each subject, we evaluated the decoders 100 times with shuffled labels to create per-subject chance-level results. Then, we randomly selected one of the 100 chance-level results for each of the 6 subjects and calculated group-level statistics (TFCE values) the exact same way as described in the preceding paragraph. We repeated this procedure 10,000 times resulting in 10,000 permuted group-level results. We ensured that every permutation was unique, i.e. no two permutations were based on the same combination of selected chance-level results. Based on this null distribution, we calculated p-values for each vertex by calculating the proportion of

[6] Details on how this size was selected are outlined in Appendix 5.

[7] We used the default hyperparameters of $h = 2$ and $e = 1$ for surface-based TFCE (***Jenkinson et al., 2012***).

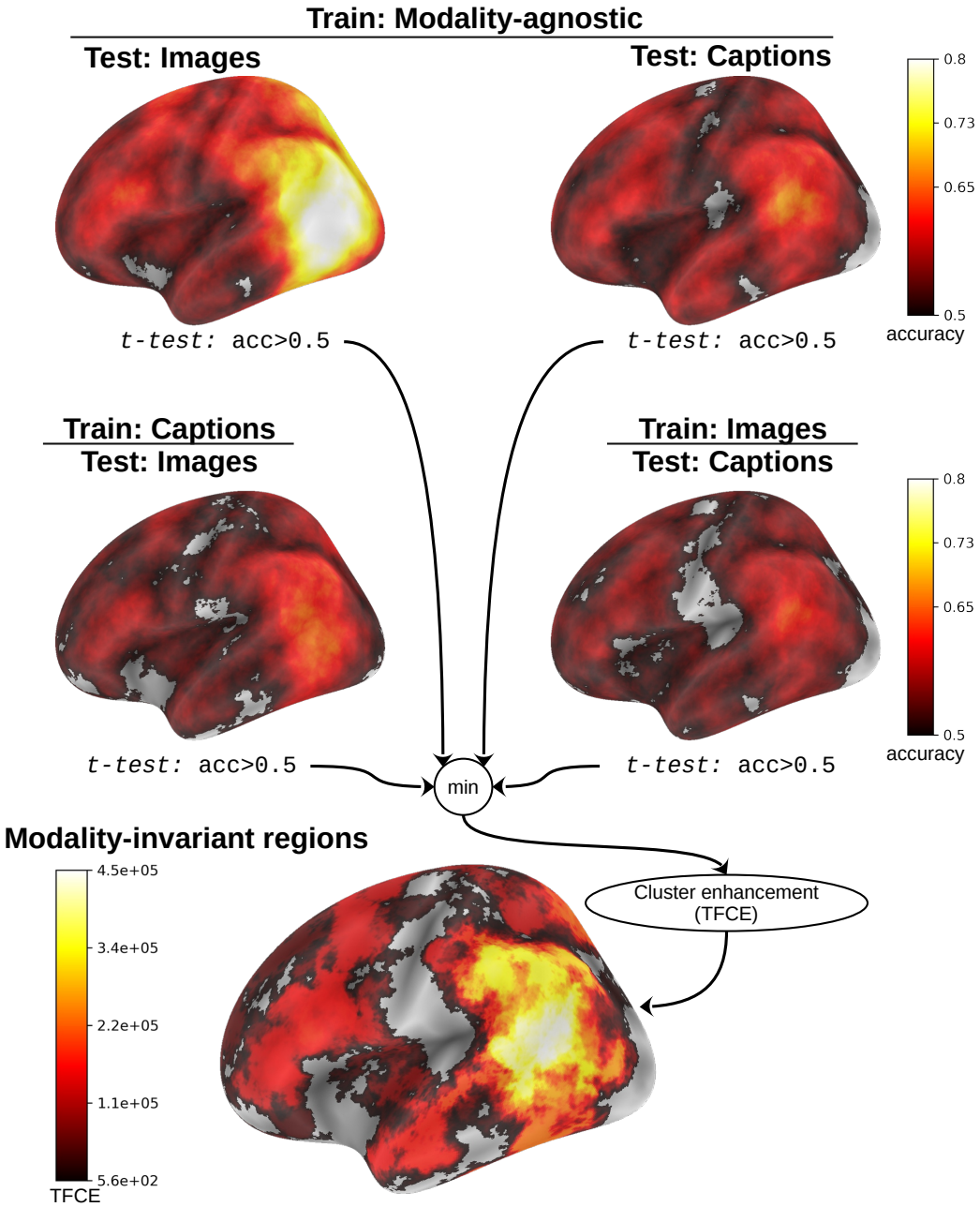


Figure 8. Searchlight method to identify modality-invariant ROIs. The top plots show performance (pairwise accuracy averaged over subjects) of modality-agnostic decoders for decoding images (top left) and decoding captions (top right). In the second row, we display cross-decoding performances: On the left, modality-specific decoders trained on captions are evaluated on images. On the right, modality-specific decoders trained on images are evaluated on captions. We identified modality-invariant ROIs as clusters in which all 4 decoding accuracies are above chance by taking the minimum of the respective t-values at each location, then performed TFCE to calculate cluster values. The plot only shows left medial views of the brain to illustrate the method, different views of all resulting clusters are shown in Figure 9.

sampled permutations where the TFCE value was greater than the observed TFCE value. To control for multiple comparisons across space, we always considered the maximum TFCE score across vertices for each group-level permutation (***Smith and Nichols, 2009***).

The results of the surface-based searchlight analysis are presented in Figure 9. The analysis re-

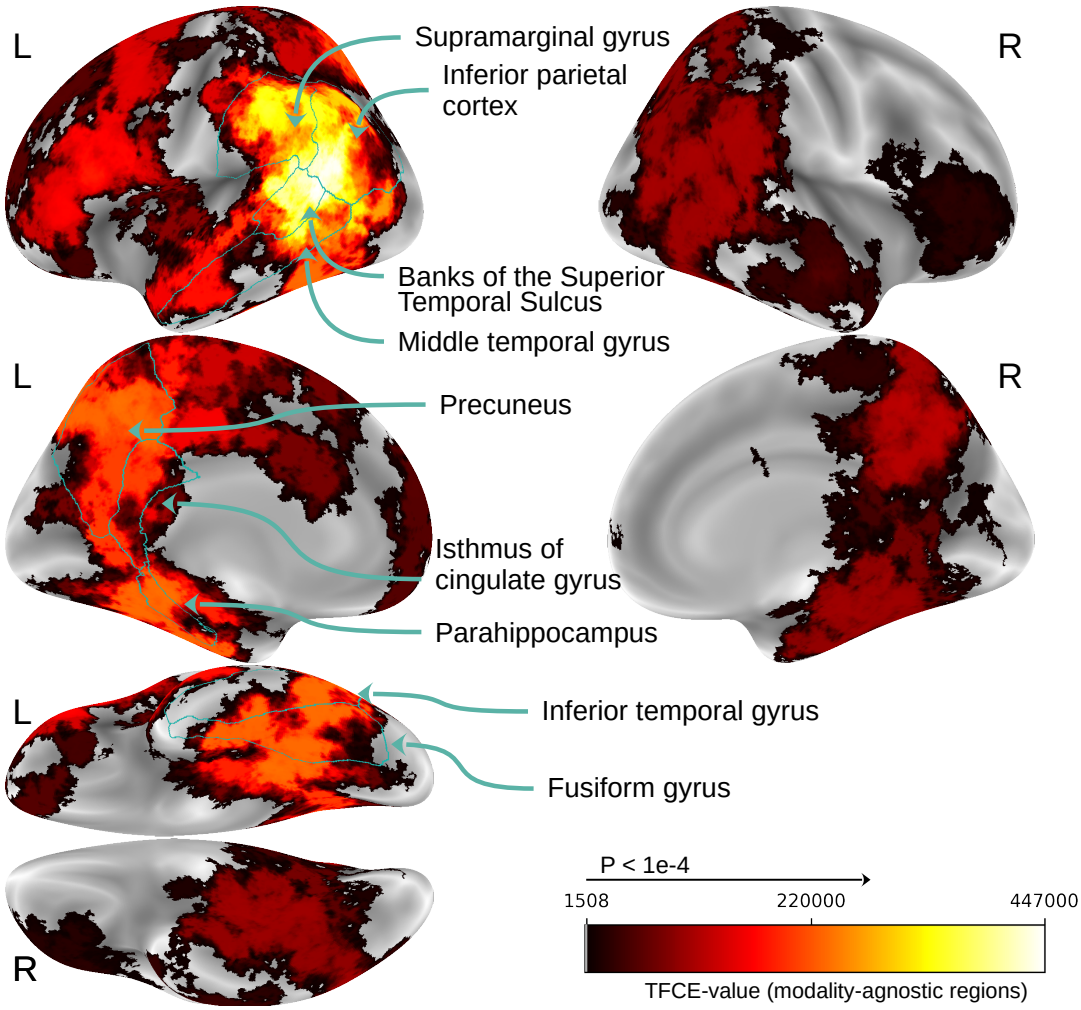


Figure 9. Searchlight results for modality-invariant regions. Maps thresholded at TFCE value of 1508, which is the significance threshold value for which $p < 10^{-4}$ based on the permutation testing. Regions with highest cluster values are outlined and annotated based on the Desikan-Killiany atlas (***Desikan et al., 2006***).

vealed that modality-invariant patterns are actually widespread across the brain, especially on the left hemisphere. Peak cluster values were found in the left supramarginal gyrus, inferior parietal gyrus and posterior superior temporal sulcus. Regions belonging to the precuneus, isthmus of the cingulate gyrus, parahippocampus, middle temporal gyrus, inferior temporal gyrus and fusiform gyrus also showed high cluster values.

Imagery Decoding

Finally, we evaluated the ability of decoders trained on the fMRI data with perceptual input to decode stimuli during the imagery conditions.

For a modality-agnostic decoder trained on the whole-brain data, the imagery pairwise decoding accuracy reaches 84.48% (averaged across subjects and model features) when using the 3 imagery stimuli as the candidate set.[8] When the whole test set is added to the candidate set (in total: 73 stimuli), the average pairwise accuracy drops to 72.47%. This substantial drop in performance is most likely explained by the fact that the predicted features for the imagery trials were standardized using only 3 stimuli, and this transformation emphasized differences that enabled

[8]Note that we used the ground-truth caption and corresponding image from COCO in this candidate set, and not the sketches drawn by the subjects.

distinguishing the 3 imagery trials but do not generalize to the larger test set.[9]

As expected, we also found that modality-agnostic decoders are better suited for imagery decoding than modality-specific decoders. We compared the imagery decoding accuracy of both decoder types taking into account the results for all features and all subjects. To this end, we performed two repeated measures ANOVAs (grouping the data by subject), once comparing the accuracy values of modality-agnostic decoders with those of modality-specific decoders trained on images, and once comparing modality-agnostic decoders to modality-specific decoders trained on captions. The average decoding accuracies were 69.42% for a modality-specific decoder trained on images and 70.02% for a modality-specific decoder trained on captions (vs. 72.47% for a modality-agnostic decoder, as mentioned above). In both comparisons, the accuracy values for the two decoder types were significantly different (when comparing modality-agnostic decoders to modality-specific decoders trained on images: decoder_type : $\beta = 0.03, SE = 0.011, p < 0.01$; and when comparing to modality-specific decoders trained on captions: decoder_type : $\beta = 0.024, SE = 0.012, p < 0.04$).

Appendix 7 presents qualitative decoding results for the imagery trials for each subject as well as the sketches of the mental images drawn at the end of the experiment. As expected, the results are worse than those for perceived stimuli, but for several subjects it was possible to decode some major semantic concepts.

We further computed the imagery decoding accuracy during the searchlight analysis. Figure 10 shows the result clusters for decoding imagery (using the whole test set + the 3 imagery trials as potential candidates). To assess statistical significance, we employed the same permutation testing procedure with bootstrapping as used for identifying modality-invariant regions (10,000 permutations based on random combinations of 100 chance distributions for each subject).

We observe that many regions that were found to contain modality-invariant patterns (cf. Figure 9) are also regions in which decoding of mental imagery is possible.

One main difference is that the imagery decoding clusters appear to be less left-lateralized than the modality-invariant region clusters (peak cluster values can be found both on the right inferior parietal cortex and bilaterally in the precuneus). To estimate overlap of the regions allowing for imagery decoding and modality-invariant regions we calculated the correlation between the TFCE values that were used for identifying modality-invariant regions (Figure 9) and the TFCE values for imagery decoding (Figure 10). The Pearson correlation score for the left hemisphere is 0.41 ($p < 1e-8$), and for the right hemisphere 0.62 ($p < 1e-8$). Importantly, these correlation scores are substantially higher when compared to the correlation with decoding accuracy of modality-specific decoders: The correlation between the TFCE values for imagery decoding and TFCE values for image decoding of a modality-specific decoder trained on images is 0.28 on the left hemisphere and 0.40 on the right hemisphere. When using TFCE values based on the caption decoding accuracy of a modality-specific decoder trained on captions we obtain 0.19 on the left hemisphere and 0.45 on the right hemisphere.

Discussion

In this work, we introduced modality-agnostic decoders, which are trained to decode stimuli from different modalities using a single model. Such modality-agnostic decoders are specifically trained to pick up on modality-invariant representations, enabling a performance increase over modality-specific decoders when decoding linguistic stimuli in the form of captions as well as when decoding mental imagery. Previous work on decoding with large-scale brain imaging datasets has largely focused on modality-specific cortices, such as the visual cortex for decoding natural scenes (***Shen et al., 2019***; ***Beliy et al., 2019***; ***Lin et al., 2022***; ***Takagi and Nishimoto, 2023***; ***Ozcelik and VanRullen, 2023***). Our work demonstrates the advantages of leveraging higher-level representations that are

[9]We also attempted decoding without standardization of the predicted feature vectors, but this led to much lower performance.

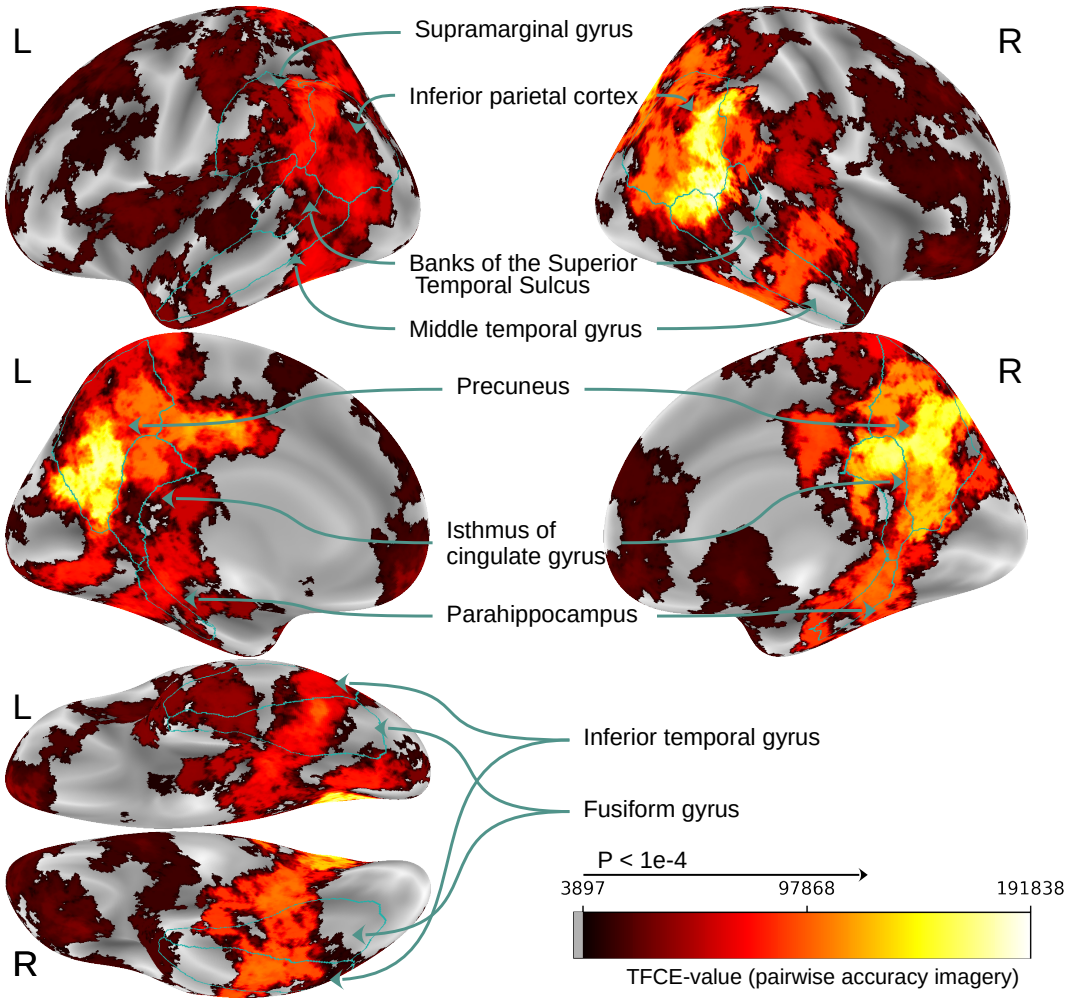


Figure 10. Searchlight results for imagery decoding. Maps thresholded at TFCE values that surpass the significance threshold of $p < 10^{-4}$ based on a permutation test. Maps thresholded at TFCE value of 3897, which is the significance threshold value for which $p < 10^{-4}$ based on the permutation testing. We used the pairwise accuracy for imagery decoding using the large candidate set of 73 stimuli. We outlined the same regions as in Figure 9 to facilitate comparison.

shared between modalities, and which we found to be present in widespread areas of the human cortex. In the following, we will further discuss the nature and location of the modality-invariant representations in light of previous experimental and theoretical work.

The results of our searchlight analysis suggest that modality-invariant representations are found in temporal, parietal, and frontal regions. Peak cluster values are found on the border of the temporal and parietal cortices on the left hemisphere (cf. Figure 9). All areas with high cluster values confirm findings from previous studies: The left precuneus (***Shinkareva et al., 2011***; ***Fairhall and Caramazza, 2013***; ***Popham et al., 2021***; ***Handjaras et al., 2016***), posterior cingulate/ retrosplenial cortex (***Fairhall and Caramazza, 2013***; ***Handjaras et al., 2016***), supramarginal gyrus (***Shinkareva et al., 2011***), inferior parietal cortex (***Man et al., 2012***; ***Vandenberghe et al., 1996***; ***Shinkareva et al., 2011***; ***Devereux et al., 2013***; ***Popham et al., 2021***; ***Simanova et al., 2014***; ***Handjaras et al., 2016***), superior temporal sulcus (***Man et al., 2012***), middle temporal gyrus (***Vandenberghe et al., 1996***; ***Shinkareva et al., 2011***; ***Fairhall and Caramazza, 2013***; ***Devereux et al., 2013***; ***Handjaras et al., 2016***), inferior temporal gyrus (***Vandenberghe et al., 1996***; ***Shinkareva et al., 2011***; ***Fairhall and Caramazza, 2013***; ***Simanova et al., 2014***; ***Handjaras et al., 2016***), fusiform gyrus (***Vandenberghe***

et al., 1996; *Moore and Price, 1999*; *Bright et al., 2004*; *Shinkareva et al., 2011*; *Fairhall and Caramazza, 2013*; *Simanova et al., 2014*), and parahippocampus (*Vandenberghe et al., 1996*). However, previous studies have led to contradicting results regarding the locality of modality-invariant regions (they were identifying varying subsets of these regions; see also Appendix 4), probably due to the limited number and artificial nature of stimuli employed. Our method identified *almost all* of the previously proposed regions as regions with modality-invariant patterns, highlighting the advantage of our searchlight decoding approach and the large multimodal dataset in which subjects are viewing photographs of complex natural scenes and reading full English sentences.[10]

According to a range of theories, modality-invariant representations are tightly linked to (lexical-) semantic representations (*Simmons and Barsalou, 2003*; *Binder et al., 2009*; *Meschke and Gallant, 2024*). Most importantly, a range of studies that aimed to identify brain regions linked to semantic/conceptual representations by asking subjects to perform tasks that require semantic processing of words found evidence for such regions that overlap to a high degree with the regions identified in our study (*Fernandino et al., 2016*; *Martin et al., 2018*; *Carota et al., 2021*; *Fernandino et al., 2022*; *Tong et al., 2022*). A strong link between these systems could also explain our result that modality-agnostic decoders based on unimodal representations from language models are performing as well as decoders based on multimodal representations (cf. Figure 4), as well as the partial left-lateralization of the identified modality-invariant regions.

The fact that the presence of modality-invariant patterns is positively correlated with the imagery decoding performance in different locations provides further evidence that the identified patterns are truly modality-invariant (they are not only shared between vision and language, but also for mental imagery). We further found that decoders trained exclusively on data for which participants were exposed to perceptual input do generalize to imagery trials, confirming previous findings that were based on more limited stimulus sets (*Stokes et al., 2009*; *Reddy et al., 2010*; *Lee et al., 2012*; *Johnson and Johnson, 2014*; *Naselaris et al., 2015*). Moreover, we found that modality-agnostic decoders outperform modality-specific decoders in terms of imagery decoding, demonstrating another advantage of this new type of decoder explicitly trained to pick up modality-invariant patterns.

The findings of our searchlight analysis for imagery decoding suggest that mental imagery indeed involves a large network of regions across both hemispheres of the cerebral cortex. This includes high-level visual areas, parietal areas such as the precuneus and inferior parietal cortex and several frontal regions, but also parts of the early visual cortex. Results are highly similar on both hemispheres, highlighting the involvement of large-scale bilateral brain networks during mental imagery of complex scenes.

While there are lesion studies on hemispheric asymmetries that suggest that regions in the left hemisphere are crucial for mental imagery (*Farah, 1984*; *Bartolomeo, 2002*), a more recent review that additionally considers evidence from neuroimaging and direct cortical stimulation studies suggests that frontoparietal networks in both hemispheres are involved in mental imagery, and that lateralization patterns can be found in the temporal lobes (*Liu et al., 2022*). Such lateralization was found to depend on the nature of the imagined items, the imagination of objects and words involving the left inferior temporal cortex while the imagination of faces and people was found to be more right-lateralized and the imagination of complex scenes (as in our study) leads to significant activity in both hemispheres (*O'Craven and Kanwisher, 2000*; *Steel et al., 2021*; *Spagna et al., 2021*). Crucially, in more recent decoding studies, results were either observed bilaterally, or the analyses did not target hemispheric asymmetries (*Reddy et al., 2010*; *Lee et al., 2012*).[11] Our work shows for the first time results of a searchlight analysis of imagery decoding of complex visual scenes.

[10] The left superior occipital gyrus was not identified in our study, but in previous studies by *Vandenberghe et al.* (*1996*); *Shinkareva et al.* (*2011*). However, we found that a major part of the left superior occipital *sulcus* represents modality-invariant information. Further, *Jung et al.* (*2018*) found modality-invariant patterns in the right superior frontal gyrus. One major difference between their study and ours is that they used auditory input as a second modality instead of text. Further work is required to investigate to what extent the modality-invariant regions identified in our work generalize to all modalities.

[11] But see *Stokes et al.* (*2009*) in which perception-to-imagery generalization of single letters was left-lateralized.

Above-chance decoding is possible on both hemispheres, with the highest decoding accuracies in the precuneus and the right inferior parietal cortex and the superior temporal sulcus. Future investigations with larger sets of imagined scenes could address the question of whether lateralization patterns depend on the nature of the imagined objects.

According to the results of our searchlight analysis, the anterior temporal lobes are not among the regions with the highest probability of being modality-invariant, contradicting the hypothesis of the hub-and-spoke theory that these areas are the major semantic hub in the brain. However, MRI signals from these regions have a lower signal-to-noise ratio with standard fMRI pulse sequences (***Devlin et al., 2000***; ***Embleton et al., 2010***). A more targeted study with an adapted fMRI protocol would be required to shed light on the nature of patterns in these regions. More generally, the hub-and-spoke theory also puts emphasis on the role of spokes for the formation of conceptual representations (***Pobric et al., 2010***; ***Ralph et al., 2017***). Future work could be aimed at testing the hub-and-spoke theory proposal that features in hierarchically lower level representation spaces of the spokes are not directly relatable to features in the representation space of the hubs: Object representations in the spokes are based on interpretable features (e.g. shape, color, affordances of an object) and get translated into another representational format that is representing conceptual similarities (but its dimensions do not directly map to interpretable features) in the semantic hub (***Frisby et al., 2023***). To test this hypothesis, modality-invariant representations in the anterior temporal lobes (measured with targeted fMRI pulse sequences) could be compared to representations in candidate regions for modality-specific spokes using RSA.

The modality-invariant regions we found in the searchlight analysis can also be seen as candidates for convergence zones, in which increasingly abstract representations are formed (***Damasio, 1989***; ***Tranel et al., 1997***; ***Meyer and Damasio, 2009***). To obtain further insight into the hierarchical organization of these zones, future work could take advantage of the improved temporal resolution of other brain imaging techniques such as MEG to explore in which areas modality-invariant patterns are formed first, and how they are being transformed when spreading to higher-level areas of the brain (***Dirani and Pylkkänen, 2024***; ***Benchetrit et al., 2024***).

In line with the GRAPES framework, (***Martin, 2009***, ***2016***), we found that modality-invariant representations are distributed across temporal and parietal areas. To test the related hypothesis that conceptual information is organized in domains, we plan to use RSA to understand which kind of semantic information is represented in the different modality-invariant regions identified.

Finally, our results can be interpreted with respect to the Global Workspace Theory. All modality-invariant regions are good candidate regions for a global workspace. They could, however, also be part of modality-specific modules that get activated in a modality-invariant fashion through a "broadcast" operation as a stimulus is perceived consciously (***Baars, 1993***, ***2005***). To distinguish these two cases, an experimental manipulation of attention could be used: according to Global Workspace Theory, attention is required for information to enter the workspace, but not for the workspace signals to reach other brain regions via broadcast. In the future, we plan to investigate how modality-invariant patterns are modulated by attention, by analyzing additional test sessions from the same subjects in which they were instructed in specific runs to pay attention to only one of the modalities.[12]

To conclude, the results from our searchlight analysis so far are in line with all major theories on modality-invariant representations that were considered. As the dataset that this study was based on will be shared publicly, more targeted investigations can be performed by the research community in order to adjudicate between different theories.

Limitations

There are some important limitations of this study, which are mainly consequences of some design choices to enable large-scale data collection. First, the fMRI experiment relied on a one-back

[12]These sessions will be released as part of another future dataset and publication.

task, which required subjects to remember preceding stimuli, and therefore, brain activity in each trial could be influenced by the previous stimuli kept in working memory. Such unrelated activity could have impaired the performance of decoders, which were only trained to decode the presented stimulus. However, this limitation could also represent an opportunity for future work on modality-invariant representations in working memory, for example by training decoders to decode the preceding stimuli from such working memory activity. Second, the inclusion of the cross-modal one-back task could have incentivized subjects to perform visual imagery when reading the captions, or to recall descriptive words (and/or have inner speech) when viewing images. Such behavior might have facilitated the spread of modality-invariant information in the cortex. It remains an open question to what degree such widespread modality-invariant information is also present during isolated reading and visual inspection of images, without any cross-modal task.

Finally, it remains an open question whether the activation patterns in the modality-invariant regions identified in our study relate to abstract concepts or to lower-level features that are shared between the two modalities (see also Section Modality-invariant representations). Our current study did not aim to tackle this question, as it is not crucial for the goal of building modality-agnostic decoders. In the future, techniques such as RSA and encoding models based on carefully designed feature spaces could be used to explore the nature of the representations in more detail. One major challenge to overcome will be the spatial overlap of modality-dependent, modality-invariant and abstract information in brain representations (***Fernandino et al., 2016***; ***Liuzzi et al., 2020***; ***Dirani and Pylkkänen, 2024***). More generally, ***Binder*** (***2016***) puts the dichotomy between modality-invariant and abstract representations into question, considering that "there is no absolute demarcation between embodied/perceptual and abstract/conceptual representation in the brain." (p. 1098). The author argues for a hierarchical system in which representational patterns become increasingly abstract, creating a continuum from actual experiential information up to higher-level conceptual information (see also ***Andrews et al., 2014***).

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.

We thank the Inserm/UPS UMR1214 Technical Platform for their help in setting up and for the acquisitions of the MRI sequences.

References

Albers A, Kok P, Toni I, Dijkerman HC, de Lange F. Shared Representations for Working Memory and Mental Imagery in Early Visual Cortex. Current Biology. 2013 Aug; 23(15):1427–1431. https://linkinghub.elsevier.com/retrieve/pii/S0960982213006908, doi: 10.1016/j.cub.2013.05.065.

Aliko S, Huang J, Gheorghiu F, Meliss S, Skipper JI. A naturalistic neuroimaging database for understanding the brain using ecological stimuli. Scientific Data. 2020 Oct; 7(1):347. doi: 10.1038/s41597-020-00680-2.

Allen EJ, St-Yves G, Wu Y, Breedlove JL, Prince JS, Dowdle LT, Nau M, Caron B, Pestilli F, Charest I, Hutchinson JB, Naselaris T, Kay K. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. Nature Neuroscience. 2022; 25(1):116–126. doi: 10.1038/s41593-021-00962-x.

Andrews M, Frank S, Vigliocco G. Reconciling Embodied and Distributional Accounts of Meaning in Language. Topics in Cognitive Science. 2014 Jul; 6(3):359–370. https://onlinelibrary.wiley.com/doi/10.1111/tops.12096, doi: 10.1111/tops.12096.

Ashburner J, Barnes G, Chen CC, Daunizeau J, Flandin G, Friston K, Kiebel S, Kilner J, Litvak V, Moran R. SPM12. Wellcome Trust Centre for Neuroimaging. 2014; .

Baars BJ. A cognitive theory of consciousness. Cambridge University Press; 1993.

Baars BJ. Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. In: *Progress in Brain Research*, vol. 150 Elsevier; 2005.p. 45–53. doi: 10.1016/S0079-6123(05)50004-9.

Barsalou LW. On Staying Grounded and Avoiding Quixotic Dead Ends. Psychonomic Bulletin & Review. 2016; 23(4):1122–1142.

Bartolomeo P. The Relationship Between Visual Perception and Visual Mental Imagery: A Reappraisal of the Neuropsychological Evidence. Cortex. 2002 Jan; 38(3):357–378. https://linkinghub.elsevier.com/retrieve/pii/S0010945208706658, doi: 10.1016/S0010-9452(08)70665-8.

Beliy R, Gaziv G, Hoogi A, Strappini F, Golan T, Irani M. From voxels to pixels and back: Self-supervision in natural-image reconstruction from fMRI. In: *NeurIPS*; 2019. .

Benchetrit Y, Banville H, King JR. Brain decoding: toward real-time reconstruction of visual perception. In: *The Twelfth International Conference on Learning Representations*; 2024. .

Binder JR. In defense of abstract conceptual representations. Psychonomic Bulletin & Review. 2016 Aug; 23(4):1096–1108. http://link.springer.com/10.3758/s13423-015-0909-1, doi: 10.3758/s13423-015-0909-1.

Binder JR, Desai RH, Graves WW, Conant LL. Where Is the Semantic System? A Critical Review and Meta-Analysis of 120 Functional Neuroimaging Studies. Cerebral Cortex. 2009; 19(12):2767–2796.

Boccia M, Sulpizio V, Teghil A, Palermo L, Piccardi L, Galati G, Guariglia C. The dynamic contribution of the high-level visual cortex to imagery and perception. Human Brain Mapping. 2019 Jan; 40(8):2449–2463. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6865452/, doi: 10.1002/hbm.24535.

Boyle JA, Pinsard B, Boukhdhir A, Belleville S, Brambatti S, Chen J, Cohen-Adad J, Cyr A, Fuente Rainville P, Bellec P. The Courtois project on neuronal modelling-first data release. In: *26th annual meeting of the organization for human brain mapping*; 2020. .

Brennan JR. Language and the brain: a slim guide to neurolinguistics. Oxford University Press; 2022. https://books.google.com/books?hl=en&lr=&id=cDRtEAAAQBAJ&oi=fnd&pg=PP1&dq=brennan+language++brain&ots=9SYgp7FBJy&sig=PTuVqNSsF7NlwWWNboXe-ruH7uM.

Brennan JR, Hale JT. Hierarchical structure guides rapid linguistic predictions during naturalistic listening. PLOS ONE. 2019 Jan; 14(1):e0207741. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0207741, doi: 10.1371/journal.pone.0207741, publisher: Public Library of Science.

Bright P, Moss H, Tyler LK. Unitary vs multiple semantics: PET studies of word and picture processing. Brain and Language. 2004; 89(3):417–432.

Carota F, Nili H, Pulvermüller F, Kriegeskorte N. Distinct fronto-temporal substrates of distributional and taxonomic similarity among words: evidence from RSA of BOLD signals. NeuroImage. 2021; 224:117408.

Chang N, Pyles JA, Marcus A, Gupta A, Tarr MJ, Aminoff EM. BOLD5000, a public fMRI dataset while viewing 5000 visual images. Scientific Data. 2019; 6(1):49. doi: 10.1038/s41597-019-0052-3.

Cox DD, Savoy RL. Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. NeuroImage. 2003 Jun; 19(2):261–270. https://www.sciencedirect.com/science/article/pii/S1053811903000491, doi: 10.1016/S1053-8119(03)00049-1.

Damasio AR. The Brain Binds Entities and Events by Multiregional Activation from Convergence Zones. Neural Computation. 1989; 1(1):123–132.

Damasio H, Tranel D, Grabowski T, Adolphs R, Damasio A. Neural systems behind word and concept retrieval. Cognition. 2004 May; 92(1-2):179–229. https://linkinghub.elsevier.com/retrieve/pii/S0010027703002312, doi: 10.1016/j.cognition.2002.07.001.

Desikan RS, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, Buckner RL, Dale AM, Maguire RP, Hyman BT, Albert MS, Killiany RJ. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. NeuroImage. 2006; 31(3):968–980.

Devereux BJ, Clarke A, Marouchos A, Tyler LK. Representational Similarity Analysis Reveals Commonalities and Differences in the Semantic Processing of Words and Objects. The Journal of Neuroscience. 2013; 33(48).

Devlin J, Chang MW, Lee K, Toutanova K. 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)* Minneapolis, Minnesota: Association for Computational Linguistics; 2019. p. 4171–4186. https://aclanthology.org/N19-1423, doi: 10.18653/v1/N19-1423.

Devlin JT, Russell RP, Davis MH, Price CJ, Wilson J, Moss HE, Matthews PM, Tyler LK. Susceptibility-Induced Loss of Signal: Comparing PET and fMRI on a Semantic Task. NeuroImage. 2000 Jun; 11(6):589–600. https://linkinghub.elsevier.com/retrieve/pii/S1053811900905950, doi: 10.1006/nimg.2000.0595.

Dijkstra N, Bosch SE, Van Gerven MAJ. Shared Neural Mechanisms of Visual Perception and Imagery. Trends in Cognitive Sciences. 2019 May; 23(5):423–434. https://linkinghub.elsevier.com/retrieve/pii/S1364661319300592, doi: 10.1016/j.tics.2019.02.004.

Dirani J, Pylkkänen L. MEG Evidence That Modality-Independent Conceptual Representations Contain Semantic and Visual Features. Journal of Neuroscience. 2024 Jul; 44(27). https://www.jneurosci.org/content/44/27/e0326242024, doi: 10.1523/JNEUROSCI.0326-24.2024, publisher: Society for Neuroscience Section: Research Articles.

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: *International Conference on Learning Representations*; 2020. .

Embleton KV, Haroon HA, Morris DM, Ralph MAL, Parker GJM. Distortion correction for diffusion-weighted MRI tractography and fMRI in the temporal lobes. Human Brain Mapping. 2010; 31(10):1570–1587. https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.20959, doi: 10.1002/hbm.20959, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hbm.20959.

Fairhall SL, Caramazza A. Brain Regions That Represent Amodal Conceptual Knowledge. Journal of Neuroscience. 2013; 33(25):10552–10558.

Farah MJ. The neurological basis of mental imagery: A componential analysis. Cognition. 1984 Dec; 18(1-3):245–272. https://linkinghub.elsevier.com/retrieve/pii/001002778490026X, doi: 10.1016/0010-0277(84)90026-X.

Fedorenko E, Behr MK, Kanwisher N. Functional specificity for high-level linguistic processing in the human brain. Proceedings of the National Academy of Sciences. 2011 Sep; 108(39):16428–16433. https://www.pnas.org/doi/10.1073/pnas.1112937108, doi: 10.1073/pnas.1112937108, publisher: Proceedings of the National Academy of Sciences.

Fedorenko E, Hsieh PJ, Nieto-Castañón A, Whitfield-Gabrieli S, Kanwisher N. New Method for fMRI Investigations of Language: Defining ROIs Functionally in Individual Subjects. Journal of Neurophysiology. 2010 Aug; 104(2):1177–1194. https://www.physiology.org/doi/10.1152/jn.00032.2010, doi: 10.1152/jn.00032.2010.

Felleman DJ, Van Essen DC. Distributed hierarchical processing in the primate cerebral cortex. Cerebral cortex (New York, NY. 1991 Jan; 1(1):1–47. https://doi.org/10.1093/cercor/1.1.1-a, doi: 10.1093/cercor/1.1.1-a.

Fernandino L, Binder JR, Desai RH, Pendl SL, Humphries CJ, Gross WL, Conant LL, Seidenberg MS. Concept Representation Reflects Multimodal Abstraction: A Framework for Embodied Semantics. Cerebral Cortex. 2016; 26(5):2018–2034.

Fernandino L, Tong JQ, Conant LL, Humphries CJ, Binder JR. Decoding the information structure underlying the neural representation of concepts. Proceedings of the National Academy of Sciences. 2022; 119(6).

Fischl B. FreeSurfer. NeuroImage. 2012; 62(2):774–781. doi: 10.1016/j.neuroimage.2012.01.021, place: Netherlands Publisher: Elsevier Science.

Friederici AD. Language in Our Brain: The Origins of a Uniquely Human Capacity. The MIT Press; 2017. https://direct.mit.edu/books/oa-monograph/3653/Language-in-Our-BrainThe-Origins-of-a-Uniquely, doi: 10.7551/mitpress/11173.001.0001.

Frisby SL, Halai AD, Cox CR, Lambon Ralph MA, Rogers TT. Decoding semantic representations in mind and brain. Trends in Cognitive Sciences. 2023 Mar; 27(3):258–281. https://linkinghub.elsevier.com/retrieve/pii/S1364661322003230, doi: 10.1016/j.tics.2022.12.006.

Gainotti G. What the locus of brain lesion tells us about the nature of the cognitive defect underlying category-specific disorders: a review. Cortex; a Journal Devoted to the Study of the Nervous System and Behavior. 2000 Sep; 36(4):539–559. doi: 10.1016/s0010-9452(08)70537-9.

Girdhar R, El-Nouby A, Liu Z, Singh M, Alwala KV, Joulin A, Misra I. ImageBind: One Embedding Space To Bind Them All. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; 2023. p. 15180–15190.

Gorgolewski K, Burns CD, Madison C, Clark D, Halchenko YO, Waskom ML, Ghosh SS. Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. Frontiers in neuroinformatics. 2011; 5:12318. Publisher: Frontiers.

Grill-Spector K, Malach R. The Human Visual Cortex. Annual Review of Neuroscience. 2004 Jul; 27(1):649–677. https://www.annualreviews.org/doi/10.1146/annurev.neuro.27.070203.144220, doi: 10.1146/annurev.neuro.27.070203.144220.

Hajhajate D, Kaufmann BC, Liu J, Siuda-Krzywicka K, Bartolomeo P. The connectional anatomy of visual mental imagery: evidence from a patient with left occipito-temporal damage. Brain Structure & Function. 2022 Dec; 227(9):3075–3083. doi: 10.1007/s00429-022-02505-x.

Handjaras G, Ricciardi E, Leo A, Lenci A, Cecchetti L, Cosottini M, Marotta G, Pietrini P. How concepts are encoded in the human brain: A modality independent, category-based cortical organization of semantic knowledge. NeuroImage. 2016; 135:232–242.

Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P. Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. Science. 2001 Sep; 293(5539):2425–2430. https://www.science.org/doi/10.1126/science.1063736, doi: 10.1126/science.1063736.

He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2016. p. 770–778.

Huth AG, Nishimoto S, Vu AT, Gallant JL. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron. 2012 Dec; 76(6):1210–1224. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3556488/, doi: 10.1016/j.neuron.2012.10.014.

Jefferies E, Patterson K, Jones RW, Ralph MAL. Comprehension of concrete and abstract words in semantic dementia. Neuropsychology. 2009 Jul; 23(4):492. https://pmc.ncbi.nlm.nih.gov/articles/PMC2801065/, doi: 10.1037/a0015452.

Jenkinson M, Beckmann CF, Behrens TEJ, Woolrich MW, Smith SM. FSL. NeuroImage. 2012 Aug; 62(2):782–790. https://www.sciencedirect.com/science/article/pii/S1053811911010603, doi: 10.1016/j.neuroimage.2011.09.015.

Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Chaplot DS, Casas Ddl, Bressand F, Lengyel G, Lample G, Saulnier L, Lavaud LR, Lachaux MA, Stock P, Scao TL, Lavril T, Wang T, Lacroix T, Sayed WE, Mistral 7B. arXiv; 2023. http://arxiv.org/abs/2310.06825, doi: 10.48550/arXiv.2310.06825, arXiv:2310.06825 [cs].

Johnson MR, Johnson MK. Decoding individual natural scene representations during perception and imagery. Frontiers in Human Neuroscience. 2014 Feb; 8. https://www.frontiersin.org/journals/human-neuroscience/articles/10.3389/fnhum.2014.00059/full, doi: 10.3389/fnhum.2014.00059, publisher: Frontiers.

Jung Y, Larsen B, Walther DB. Modality-Independent Coding of Scene Categories in Prefrontal Cortex. Journal of Neuroscience. 2018 Jun; 38(26):5969–5981.

Kanwisher N. Functional specificity in the human brain: A window into the functional architecture of the mind. Proceedings of the National Academy of Sciences. 2010 Jun; 107(25):11163–11170. https://www.pnas.org/doi/abs/10.1073/pnas.1005062107, doi: 10.1073/pnas.1005062107, publisher: Proceedings of the National Academy of Sciences.

Kay KN, Naselaris T, Prenger RJ, Gallant JL. Identifying natural images from human brain activity. Nature. 2008 Mar; 452(7185):352–355. https://www.nature.com/articles/nature06713, doi: 10.1038/nature06713.

Kiefer M, Pulvermüller F. Conceptual representations in mind and brain: Theoretical developments, current evidence and future directions. Cortex. 2012 Jul; 48(7):805–825. https://linkinghub.elsevier.com/retrieve/pii/S0010945211001018, doi: 10.1016/j.cortex.2011.04.006.

Kim W, Son B, Kim I. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In: *Proceedings of the 38th International Conference on Machine Learning* PMLR; 2021. p. 5583–5594. https://proceedings.mlr.press/v139/kim21k.html, iSSN: 2640-3498.

Kosslyn SM, Pascual-Leone A, Felician O, Camposano S, Keenan JP, L W, Thompson, Ganis G, Sukel KE, Alpert NM. The Role of Area 17 in Visual Imagery: Convergent Evidence from PET and rTMS. Science. 1999 Apr; 284(5411):167–170. https://www.science.org/doi/10.1126/science.284.5411.167, doi: 10.1126/science.284.5411.167.

Kosslyn SM, Thompson WL. When is early visual cortex activated during visual mental imagery? Psychological Bulletin. 2003; 129(5):723–746. https://doi.apa.org/doi/10.1037/0033-2909.129.5.723, doi: 10.1037/0033-2909.129.5.723.

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

Kriegeskorte N, Mur M, Bandettini PA. Representational similarity analysis - connecting the branches of systems neuroscience. Frontiers in Systems Neuroscience. 2008 Nov; 2. https://www.frontiersin.org/journals/systems-neuroscience/articles/10.3389/neuro.06.004.2008/full, doi: 10.3389/neuro.06.004.2008, publisher: Frontiers.

Lambon Ralph MA, Lowe C, Rogers TT. Neural basis of category-specific semantic deficits for living things: evidence from semantic dementia, HSVE and a neural network model. Brain. 2006 Nov; 130(4):1127–1137. https://academic.oup.com/brain/article-lookup/doi/10.1093/brain/awm025, doi: 10.1093/brain/awm025.

Lee SH, Kravitz DJ, Baker CI. Disentangling visual imagery and perception of real-world objects. NeuroImage. 2012 Feb; 59(4):4064–4073. https://www.sciencedirect.com/science/article/pii/S1053811911012195, doi: 10.1016/j.neuroimage.2011.10.055.

Li J, Li D, Savarese S, Hoi S. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In: *International Conference on Machine Learning*; 2023. .

Li LH, Yatskar M, Yin D, Hsieh CJ, Chang KW. VisualBERT: A Simple and Performant Baseline for Vision and Language. arXiv:190803557 [cs]. 2019 Aug; http://arxiv.org/abs/1908.03557, arXiv: 1908.03557.

Lin S, Sprague T, Singh AK. Mind Reader: Reconstructing complex images from brain activities. Advances in Neural Information Processing Systems. 2022 Dec; 35:29624–29636.

Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft COCO: Common Objects in Context. In: *Computer Vision – ECCV 2014*, vol. 8693; 2014.p. 740–755.

Liu J, Spagna A, Bartolomeo P. Hemispheric asymmetries in visual mental imagery. Brain Structure & Function. 2022 Mar; 227(2):697–708. doi: 10.1007/s00429-021-02277-w.

Liu J, Zhan M, Hajhajate D, Spagna A, Dehaene S, Cohen L, Bartolomeo P. Visual mental imagery in typical imagers and in aphantasia: A millimeter-scale 7-T fMRI study. Cortex. 2025 Apr; 185:113–132. https://www.sciencedirect.com/science/article/pii/S0010945225000474, doi: 10.1016/j.cortex.2025.01.013.

Liuzzi AG, Aglinskas A, Fairhall SL. General and feature-based semantic representations in the semantic network. Scientific Reports. 2020 Jun; 10(1):8931. doi: 10.1038/s41598-020-65906-0.

Liuzzi AG, Bruffaerts R, Peeters R, Adamczuk K, Keuleers E, De Deyne S, Storms G, Dupont P, Vandenberghe R. Cross-modal representation of spoken and written word meaning in left pars triangularis. NeuroImage. 2017 Apr; 150:292–307. https://linkinghub.elsevier.com/retrieve/pii/S105381191730143X, doi: 10.1016/j.neuroimage.2017.02.032.

Man K, Kaplan JT, Damasio A, Meyer K. Sight and Sound Converge to Form Modality-Invariant Representations in Temporoparietal Cortex. Journal of Neuroscience. 2012; 32(47):16629–16636.

Martin A. Circuits in Mind: The Neural Foundations for Object Concepts. In: Gazzaniga MS, editor. *The Cognitive Neurosciences*, 4 ed. The MIT Press; 2009.https://direct.mit.edu/books/book/5453/chapter/3965022/Circuits-in-Mind-The-Neural-Foundations-for-Object, doi: 10.7551/mitpress/8029.003.0091.

Martin A. GRAPES—Grounding representations in action, perception, and emotion systems: How object properties and categories are represented in the human brain. Psychonomic Bulletin & Review. 2016; 23(4):979–990.

Martin CB, Douglas D, Newsome RN, Man LL, Barense MD. Integrative and distinctive coding of visual and conceptual object features in the ventral visual stream. eLife. 2018; 7.

Meschke E, Gallant J. Mapping Multimodal Conceptual Representations within the Lexical-Semantic Brain System. In: *CCN*; 2024. .

Meyer K, Damasio A. Convergence and divergence in a neural architecture for recognition and memory. Trends in Neurosciences. 2009 Jul; 32(7):376–382. https://www.cell.com/trends/neurosciences/abstract/S0166-2236(09)00090-3, doi: 10.1016/j.tins.2009.04.002, publisher: Elsevier.

Moore CJ, Price CJ. Three Distinct Ventral Occipitotemporal Regions for Reading and Object Naming. NeuroImage. 1999; 10(2):181–192.

Naselaris T, Kay KN, Nishimoto S, Gallant JL. Encoding and decoding in fMRI. NeuroImage. 2011 May; 56(2):400–410. https://www.sciencedirect.com/science/article/pii/S1053811910010657, doi: 10.1016/j.neuroimage.2010.07.073.

Naselaris T, Olman CA, Stansbury DE, Ugurbil K, Gallant JL. A voxel-wise encoding model for early visual areas decodes mental images of remembered scenes. NeuroImage. 2015 Jan; 105:215–228. https://www.sciencedirect.com/science/article/pii/S1053811914008428, doi: 10.1016/j.neuroimage.2014.10.018.

Naselaris T, Prenger RJ, Kay KN, Oliver M, Gallant JL. Bayesian Reconstruction of Natural Images from Human Brain Activity. Neuron. 2009 Sep; 63(6):902–915. https://linkinghub.elsevier.com/retrieve/pii/S0896627309006850, doi: 10.1016/j.neuron.2009.09.006.

Nastase SA, Liu YF, Hillman H, Zadbood A, Hasenfratz L, Keshavarzian N, Chen J, Honey CJ, Yeshurun Y, Regev M, Nguyen M, Chang CHC, Baldassano C, Lositsky O, Simony E, Chow MA, Leong YC, Brooks PP, Micciche E, Choe G, et al. The "Narratives" fMRI dataset for evaluating models of naturalistic language comprehension. Scientific Data. 2021 Sep; 8(1):250. https://www.nature.com/articles/s41597-021-01033-3, doi: 10.1038/s41597-021-01033-3.

Nishimoto S, Vu AT, Naselaris T, Benjamini Y, Yu B, Gallant JL. Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies. Current Biology. 2011 Oct; 21(19):1641–1646. https://www.cell.com/current-biology/abstract/S0960-9822(11)00937-7, doi: 10.1016/j.cub.2011.08.031, publisher: Elsevier.

O'Craven KM, Kanwisher N. Mental Imagery of Faces and Places Activates Corresponding Stimulus-Specific Brain Regions. Journal of Cognitive Neuroscience. 2000 Nov; 12(6):1013–1023. https://direct.mit.edu/jocn/article/12/6/1013/3493/Mental-Imagery-of-Faces-and-Places-Activates, doi: 10.1162/08989290051137549.

Visconti di Oleggio Castello M, Chauhan V, Jiahui G, Gobbini MI. An fMRI dataset in response to "The Grand Budapest Hotel", a socially-rich, naturalistic movie. Scientific Data. 2020; 7(1):383. doi: 10.1038/s41597-020-00735-4.

Oquab M, Darcet T, Moutakanni T, Vo H, Szafraniec M, Khalidov V, Fernandez P, Haziza D, Massa F, El-Nouby A, Assran M, Ballas N, Galuba W, Howes R, Huang PY, Li SW, Misra I, Rabbat M, Sharma V, Synnaeve G, et al., DINOv2: Learning Robust Visual Features without Supervision. arXiv; 2023. http://arxiv.org/abs/2304.07193, arXiv:2304.07193 [cs].

Ozcelik F, VanRullen R, Natural scene reconstruction from fMRI signals using generative latent diffusion. arXiv; 2023. http://arxiv.org/abs/2303.05334, arXiv:2303.05334 [cs, q-bio].

Patterson K, Lambon Ralph MA. The Hub-and-Spoke Hypothesis of Semantic Memory. In: *Neurobiology of Language* Elsevier; 2016.p. 765–775. https://linkinghub.elsevier.com/retrieve/pii/B9780124077942000614, doi: 10.1016/B978-0-12-407794-2.00061-4.

Pearson J. The human imagination: the cognitive neuroscience of visual mental imagery. Nature Reviews Neuroscience. 2019 Oct; 20(10):624–634. https://www.nature.com/articles/s41583-019-0202-9, doi: 10.1038/s41583-019-0202-9.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011; 12:2825–2830.

Pereira F, Lou B, Pritchett B, Ritter S, Gershman SJ, Kanwisher N, Botvinick M, Fedorenko E. Toward a universal decoder of linguistic meaning from brain activation. Nature Communications. 2018 Mar; 9(1):963. https://www.nature.com/articles/s41467-018-03068-4, doi: 10.1038/s41467-018-03068-4, number: 1 Publisher: Nature Publishing Group.

Pobric G, Jefferies E, Ralph MAL. Category-Specific versus Category-General Semantic Impairment Induced by Transcranial Magnetic Stimulation. Current Biology. 2010 May; 20(10):964–968. https://www.cell.com/current-biology/abstract/S0960-9822(10)00456-2, doi: 10.1016/j.cub.2010.03.070, publisher: Elsevier.

Popham SF, Huth AG, Bilenko NY, Deniz F, Gao JS, Nunez-Elizalde AO, Gallant JL. Visual and linguistic semantic representations are aligned at the border of human visual cortex. Nature Neuroscience. 2021; 24(11):1628–1636.

Power JD, Barnes KA, Snyder AZ, Schlaggar BL, Petersen SE. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. NeuroImage. 2012 Feb; 59(3):2142–2154. https://linkinghub.elsevier.com/retrieve/pii/S1053811911011815, doi: 10.1016/j.neuroimage.2011.10.018.

Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I. Learning Transferable Visual Models From Natural Language Supervision. Proceedings of the 38th International Conference on Machine Learning. 2021; p. 16.

Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. OpenAI blog. 2019; 1(8):9.

Ralph MAL, Jefferies E, Patterson K, Rogers TT. The neural and computational bases of semantic cognition. Nature Reviews Neuroscience. 2017; 18(1):42–55.

Reddy L, Tsuchiya N, Serre T. Reading the mind's eye: decoding category information during mental imagery. NeuroImage. 2010 Apr; 50(2):818–825. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2823980/, doi: 10.1016/j.neuroimage.2009.11.084.

Reimers N, Gurevych I. 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)* Hong Kong, China: Association for Computational Linguistics; 2019. p. 3980–3990. https://www.aclweb.org/anthology/D19-1410, doi: 10.18653/v1/D19-1410.

Rogers TT, Lambon Ralph MA, Garrard P, Bozeat S, McClelland JL, Hodges JR, Patterson K. Structure and Deterioration of Semantic Memory: A Neuropsychological and Computational Investigation. Psychological Review. 2004; 111(1):205–235. https://doi.apa.org/doi/10.1037/0033-295X.111.1.205, doi: 10.1037/0033-295X.111.1.205.

Sanchez G, Hartmann T, Fuscà M, Demarchi G, Weisz N. Decoding across sensory modalities reveals common supramodal signatures of conscious perception. Proceedings of the National Academy of Sciences. 2020 Mar; 117(13):7437–7446. https://pnas.org/doi/full/10.1073/pnas.1912584117, doi: 10.1073/pnas.1912584117.

Schoffelen JM, Oostenveld R, Lam NHL, Uddén J, Hultén A, Hagoort P. A 204-subject multimodal neuroimaging dataset to study language processing. Scientific Data. 2019; 6(1):17. doi: 10.1038/s41597-019-0020-y.

Sereno MI, Dale AM, Reppas JB, Kwong KK, Belliveau JW, Brady TJ, Rosen BR, Tootell RBH. Borders of Multiple Visual Areas in Humans Revealed by Functional Magnetic Resonance Imaging. Science. 1995 May; 268(5212):889–893. https://www.science.org/doi/10.1126/science.7754376, doi: 10.1126/science.7754376.

Shen G, Horikawa T, Majima K, Kamitani Y. Deep image reconstruction from human brain activity. PLOS Computational Biology. 2019 Jan; 15(1):e1006633. https://dx.plos.org/10.1371/journal.pcbi.1006633, doi: 10.1371/journal.pcbi.1006633.

Shinkareva SV, Malave VL, Mason RA, Mitchell TM, Just MA. Commonality of neural representations of words and pictures. NeuroImage. 2011; 54(3):2418–2425.

Simanova I, Hagoort P, Oostenveld R, van Gerven MAJ. Modality-Independent Decoding of Semantic Information from the Human Brain. Cerebral Cortex. 2014; 24(2):426–434.

Simmons WK, Barsalou LW. The similarity-in-topography principle: Reconciling theories of conceptual deficits. Cognitive Neuropsychology. 2003; 20(3-6):451–486.

Singh A, Hu R, Goswami V, Couairon G, Galuba W, Rohrbach M, Kiela D. FLAVA: A Foundational Language and Vision Alignment Model. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; 2022. p. 15638–15650.

Smith SM, Nichols TE. Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. NeuroImage. 2009; 44(1):83–98.

Snowden J, Goulding PJ, Neary D. Semantic dementia: A form of circumscribed cerebral atrophy. Behavioural Neurology. 1989; 2(3):124043. https://onlinelibrary.wiley.com/doi/abs/10.1155/1989/124043, doi: 10.1155/1989/124043, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1155/1989/124043.

Spagna A, Hajhajate D, Liu J, Bartolomeo P. Visual mental imagery engages the left fusiform gyrus, but not the early visual cortex: A meta-analysis of neuroimaging evidence. Neuroscience & Biobehavioral Reviews. 2021 Mar; 122:201–217. https://www.sciencedirect.com/science/article/pii/S0149763420307041, doi: 10.1016/j.neubiorev.2020.12.029.

Steel A, Billings MM, Silson EH, Robertson CE. A network linking scene perception and spatial memory systems in posterior cerebral cortex. Nature Communications. 2021 May; 12(1):2632. https://www.nature.com/articles/s41467-021-22848-z, doi: 10.1038/s41467-021-22848-z, publisher: Nature Publishing Group.

Steiner A, Pinto AS, Tschannen M, Keysers D, Wang X, Bitton Y, Gritsenko A, Minderer M, Sherbondy A, Long S, Qin S, Ingle R, Bugliarello E, Kazemzadeh S, Mesnard T, Alabdulmohsin I, Beyer L, Zhai X, PaliGemma 2: A Family of Versatile VLMs for Transfer. arXiv; 2024. http://arxiv.org/abs/2412.03555, doi: 10.48550/arXiv.2412.03555, arXiv:2412.03555 [cs].

Stelzer J, Chen Y, Turner R. Statistical inference and multiple testing correction in classification-based multivoxel pattern analysis (MVPA): random permutations and cluster size control. NeuroImage. 2013 Jan; 65:69–82. doi: 10.1016/j.neuroimage.2012.09.063.

Stokes M, Thompson R, Cusack R, Duncan J. Top-Down Activation of Shape-Specific Population Codes in Visual Cortex during Mental Imagery. The Journal of Neuroscience. 2009 Feb; 29(5):1565–1572. https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.4657-08.2009, doi: 10.1523/JNEUROSCI.4657-08.2009.

Takagi Y, Nishimoto S. High-Resolution Image Reconstruction With Latent Diffusion Models From Human Brain Activity. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; 2023. p. 14453–14463.

Tang J, Du M, Vo VA. Brain encoding models based on multimodal transformers can transfer across language and vision. In: *NeurIPS*; 2023. .

Tang J, LeBel A, Jain S, Huth AG. Semantic reconstruction of continuous language from non-invasive brain recordings. Nature Neuroscience. 2023 May; 26(5):858–866. https://www.nature.com/articles/s41593-023-01304-9, doi: 10.1038/s41593-023-01304-9, number: 5 Publisher: Nature Publishing Group.

Tong J, Binder JR, Humphries C, Mazurchuk S, Conant LL, Fernandino L. A Distributed Network for Multimodal Experiential Representation of Concepts. Journal of Neuroscience. 2022; 42(37):7121–7130.

Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, Bashlykov N, Batra S, Bhargava P, Bhosale S, Bikel D, Blecher L, Ferrer CC, Chen M, Cucurull G, Esiobu D, Fernandes J, Fu J, Fu W, Fuller B, et al., Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv; 2023. http://arxiv.org/abs/2307.09288, doi: 10.48550/arXiv.2307.09288.

Tranel D, Damasio H, Damasio AR. A neural basis for the retrieval of conceptual knowledge. Neuropsychologia. 1997 Oct; 35(10):1319–1327. https://www.sciencedirect.com/science/article/pii/S0028393297000857, doi: 10.1016/S0028-3932(97)00085-7.

Vandenberghe R, Price C, Wise R, Josephs O, Frackowiak RS. Functional anatomy of a common semantic system for words and pictures. Nature. 1996; 383(6597).

VanRullen R, Reddy L. Reconstructing faces from fMRI patterns using deep generative neural networks. Communications Biology. 2019 May; 2(1):1–10. https://www.nature.com/articles/s42003-019-0438-y, doi: 10.1038/s42003-019-0438-y, number: 1 Publisher: Nature Publishing Group.

Warrington EK. The Selective Impairment of Semantic Memory. Quarterly Journal of Experimental Psychology. 1975 Nov; 27(4):635–657. https://journals.sagepub.com/doi/10.1080/14640747508400525, doi: 10.1080/14640747508400525.

Warrington EK, Mccarthy RA. Categories of knowledge: Further fractionations and an attempted integration. Brain: a Journal of Neurology. 1987; .

Warrington EK, Shallice T. Category Specific Semantic Impairments. Brain: a Journal of Neurology. 1984; .

Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M. Transformers: State-of-the-art natural language processing. In: *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*; 2020. p. 38–45.

Xu X, Wu C, Rosenman S, Lal V, Che W, Duan N. BridgeTower: Building Bridges between Encoders in Vision-Language Representation Learning. Proceedings of the AAAI Conference on Artificial Intelligence. 2023 Jun; 37(9):10637–10647. https://ojs.aaai.org/index.php/AAAI/article/view/26263, doi: 10.1609/aaai.v37i9.26263.

Zhai X, Mustafa B, Kolesnikov A, Beyer L. Sigmoid Loss for Language Image Pre-Training. In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* Paris, France: IEEE; 2023. p. 11941–11952. https://ieeexplore.ieee.org/document/10377550/, doi: 10.1109/ICCV51070.2023.01100.

Zola-Morgan S. Localization of Brain Function: The Legacy of Franz Joseph Gall (1758-1828). Annual Review of Neuroscience. 1995 Mar; 18(Volume 18, 1995):359–383. https://www.annualreviews.org/content/journals/10.1146/annurev.ne.18.030195.002043, doi: 10.1146/annurev.ne.18.030195.002043, publisher: Annual Reviews.

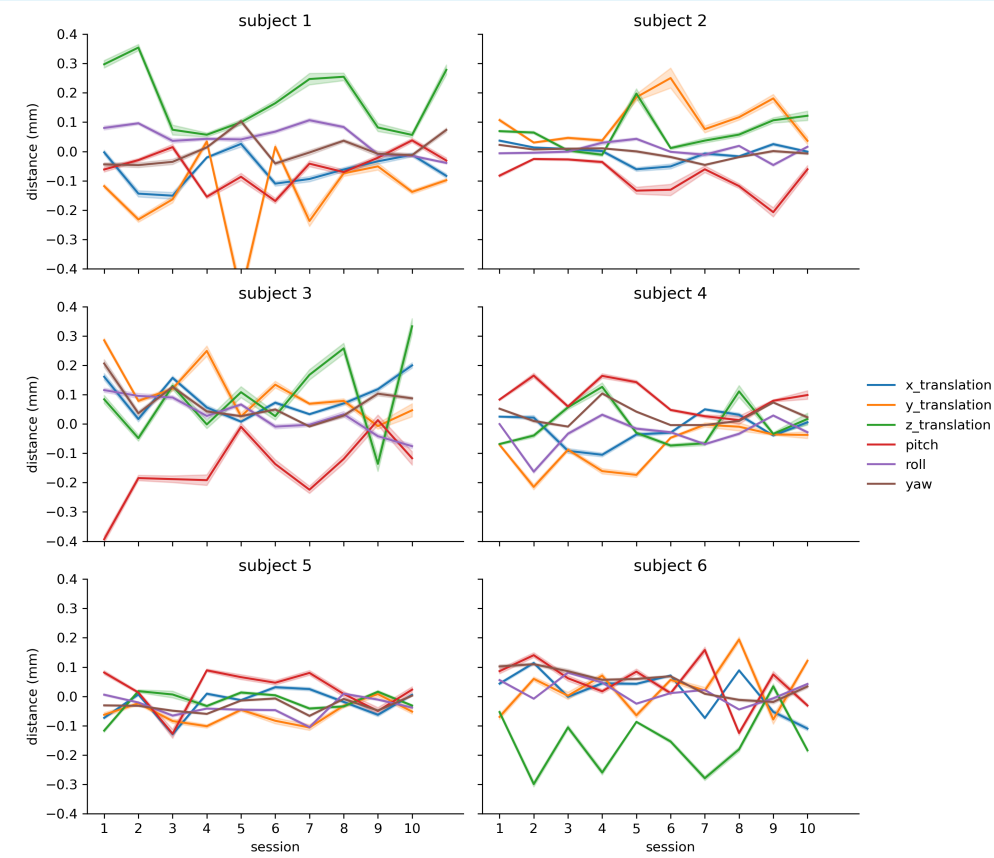
Zwaan RA. Situation models, mental simulations, and abstract concepts in discourse comprehension. Psychonomic Bulletin & Review. 2016 Aug; 23(4):1028–1034. http://link.springer.com/10.3758/s13423-015-0864-x, doi: 10.3758/s13423-015-0864-x.

Appendix 1

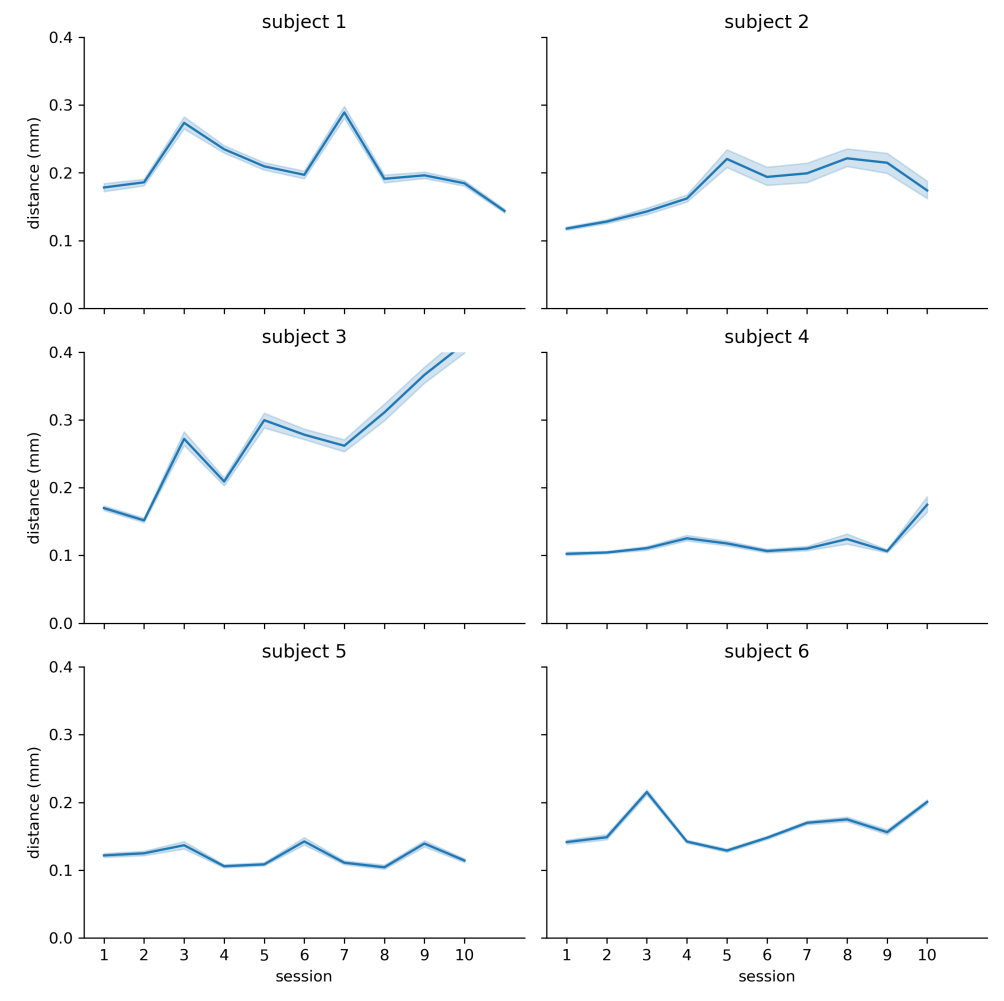
Dataset quality metrics

Head motion

Figure 1 and 2 present information on head motion for each subject. All head motion estimates are based on the realignment parameters calculated using SPM12. We found that head motion estimates are largely stable and do not vary extensively between sessions and subjects.



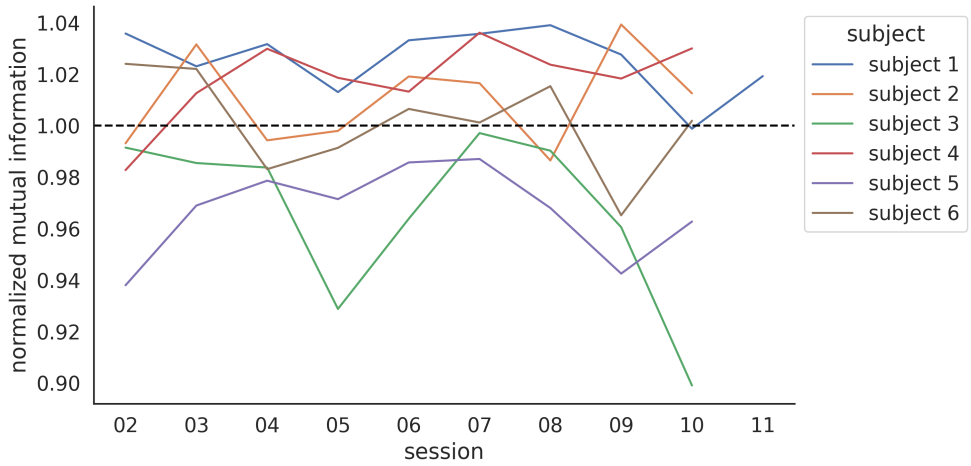
Appendix 1—figure 1. Head motion estimates for each subject. The plots show the realignment parameters as computed by SPM12 (`spm_realign`) for estimating within modality rigid body alignment. We multiplied the rotation parameters pitch, roll and yaw (originally in radian units) by 50 in order to allow interpretation in terms of millimeters of displacement for a circle of diameter 10 cm (which is approximately the mean distance from the cerebral cortex to the center of the head) (***Power et al., 2012***). The translucent error bands show 95% confidence intervals calculating using bootstrapping over all frames/runs from a session.



Appendix 1—figure 2. Framewise displacement for each subject. The measure indicates how much the head changed position from one frame to the next. We calculated framewise displacement as the sum of the absolute values of the derivatives of the six realignment parameters. The translucent error bands show 95% confidence intervals calculating using bootstrapping over all frames/runs from a session.

Intersession alignment

During preprocessing, all functional MRI data was coregistered to the anatomical scan of the first session. To assess to which degree the alignment of functional data of different sessions varies with respect to the alignment of the first session, we estimated mutual information between the coregistered scans and the reference anatomical scan. During coregistration, our pipeline based on SPM12 created a mean image of the functional data for each session to estimate transformation parameters specific for each scanning session. We calculated mutual information between the coregistered mean images and the anatomical scan of the first session. Then, we normalized all values with respect to the mutual information of the first session (i.e. the normalized mutual information value for session 1 for all subjects is 1). The results are shown in Figure 3. The plot shows that the mutual information does not deviate strongly from the baseline (minimum value: 90%) and stays in an interval of $\pm 5\%$ for most subjects.



Appendix 1—figure 3. Mutual information between the anatomical scan of the first session and functional data for each session as an indicator of intersession alignment. All values were normalized based on the mutual information from the functional data from the first session. The raw mutual information scores for the alignment during the first session are: subject 1: 0.49; subject 2: 0.55; subject 3: 0.55; subject 4: 0.53; subject 5: 0.57; subject 6: 0.57.

Appendix 2

Feature extraction details

For each target stimulus (image or caption), our database also contained an equivalent stimulus in the other modality (caption or image). In this way, we could extract model features from the corresponding image for vision models, the corresponding caption for language models, and an multimodal representation of both image and caption for the multimodal models. We used publicly available pretrained models implemented in the HuggingFace Transformers library (***Wolf et al., 2020***) or from their respective authors' repositories.

Model versions for unimodal models are as indicated in Figure 4. For multimodal models, the exact version for CLIP was `clip-vit-large-patch14`, for ViLT `vilt-b32-mlm`, for VisualBERT `visualbert-nlvr2-coco-pre`, for Imagebind `imagebind_huge`, for Bridgetower `bridgetower-large-itm-mlm-itc`, for Flava `flava-full`, for SigLip `siglip-so400m-patch14-384`, and for Paligemma2 `paligemma2-3b-pt-224`.

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

For Transformer-based vision models, we compared representations extracted by averaging the outputs for each patch with representations extracted from `[CLS]` tokens in Table 1. We found 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.

model	vision features	pairwise accuracy
dino-base	vision_features_cls	0.763
dino-base	**vision_features_mean**	**0.819**
dino-giant	vision_features_cls	0.750
dino-giant	**vision_features_mean**	**0.820**
dino-large	vision_features_cls	0.754
dino-large	**vision_features_mean**	**0.816**
vit-b-16	**vision_features_cls**	**0.788**
vit-b-16	vision_features_mean	0.772
vit-h-14	vision_features_cls	0.785
vit-h-14	**vision_features_mean**	**0.804**
vit-l-16	vision_features_cls	0.788
vit-l-16	**vision_features_mean**	**0.796**

Appendix 2—table 1. Feature comparison for vision models. Pairwise accuracy for modality-agnostic decoders based on vision features extracted by averaging the last hidden states of all patches ("vision_features_mean") compared to when using features extracted from `[CLS]` tokens ("vision_features_cls"). The method leading to the best decoding performance for each model is highlighted in bold.

For multimodal models, we also compared a range of techniques for feature extraction. The results are shown in Table 2.

model	features	vision features	language features	pairwise accuracy
blip2	**avg**	**vision_features_cls**	**lang_features_cls**	**0.851**
blip2	fused_cls	vision_features_cls	lang_features_cls	0.710
blip2	fused_mean	vision_features_cls	lang_features_cls	0.740
bridgetower	**fused_cls**			**0.815**
bridgetower	fused_mean			0.788
clip	**avg**	**vision_features_cls**	**lang_features_cls**	**0.842**
flava	**avg**	**vision_features_cls**	**lang_features_cls**	**0.842**
flava	fused_cls	vision_features_cls	lang_features_cls	0.752
flava	fused_mean	vision_features_cls	lang_features_cls	0.772
imagebind	**avg**	**vision_features_cls**	**lang_features_cls**	**0.857**
paligemma2	avg	vision_features_cls	lang_features_mean	0.829
paligemma2	**avg**	**vision_features_mean**	**lang_features_mean**	**0.848**
paligemma2	fused_mean	vision_features_mean	lang_features_mean	0.828
siglip	**avg**	**vision_features_cls**	**lang_features_cls**	**0.852**
siglip	avg	vision_features_mean	lang_features_cls	0.823
vilt	fused_cls			0.759
vilt	**fused_mean**			**0.839**
visualbert	fused_cls			0.639
visualbert	**fused_mean**			**0.743**

Appendix 2—table 2. Feature comparison for multimodal models. The features are either based on the `[CLS]` tokens from the fused representations ("fused_cls"), averaging over the fused tokens ("fused_mean"), or averaging over tokens from intermediate vision and language stream outputs ("avg"). For the last case, for some models the vision and language features can also be either based on `[CLS]` or based on averaging over all tokens/patches. The method leading to the best decoding performance for each model is highlighted in bold.

For dual-stream multimodal models, we averaged the vision and language features to create the final multimodal feature representation. For single-stream multimodal features we compared using representations extracted by averaging the outputs for each token with representations extracted from `[CLS]` token and found that the averaged output leads to better performance in almost all cases (see Table 2). Further, some single-stream models (Flava, Paligemma2 and BLIP2) allow for feature extraction based on intermediate vision and language representations in addition to a direct extraction of multimodal features (based on fused representations from the multimodal stream). We found that averaging features from these intermediate vision and language representations leads to better performance (see Table 2). For all results reported in the main paper we used the feature extraction method leading to best performance for each model.

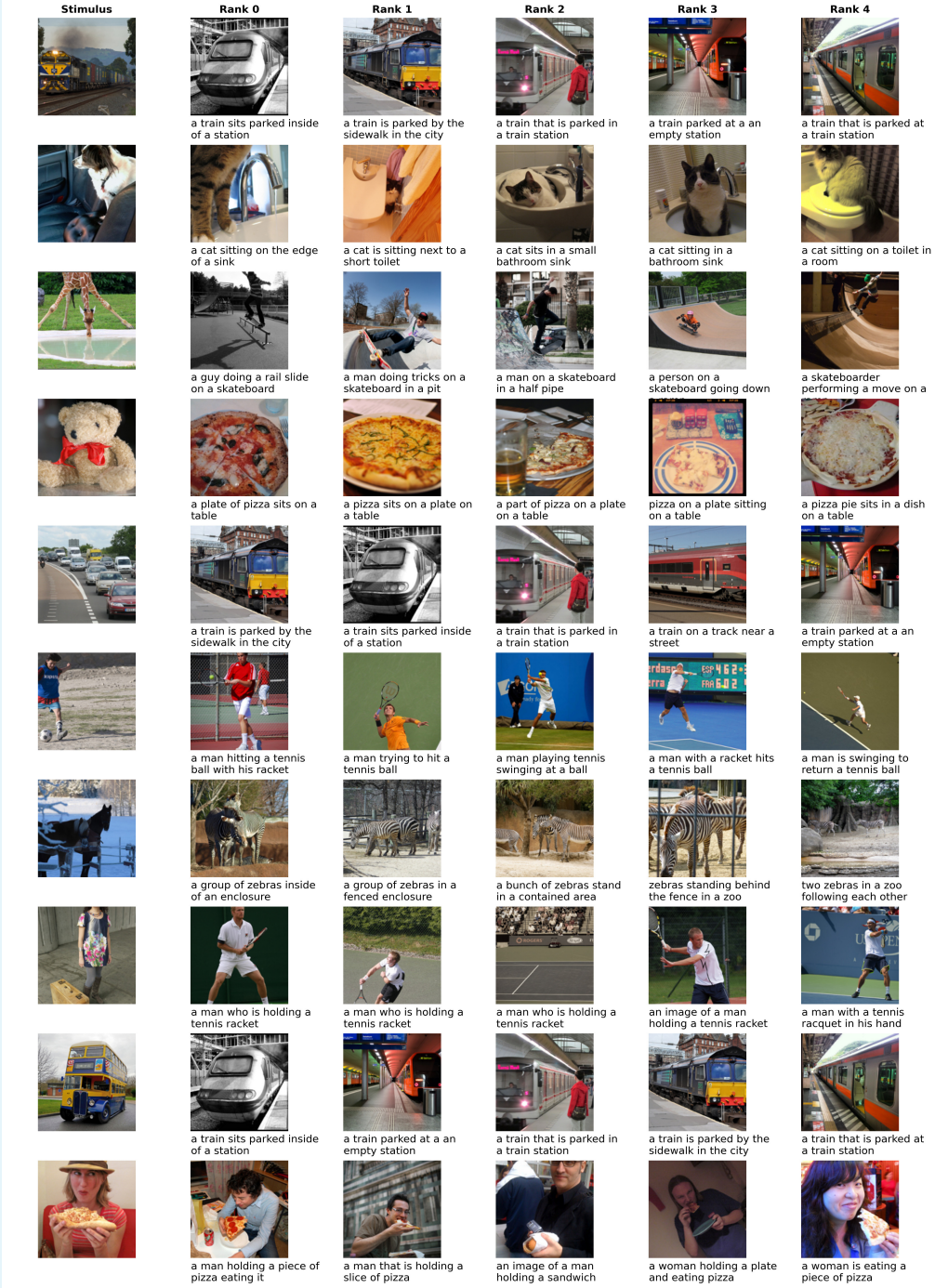
Decoder training details

The decoders were linear ridge-regression models as implemented in the scikit-learn library (***Pedregosa et al., 2011***). All training data was standardized to have mean of 0 and standard deviation of 1. The test data was standardized using the mean and standard deviation of the training data. 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.

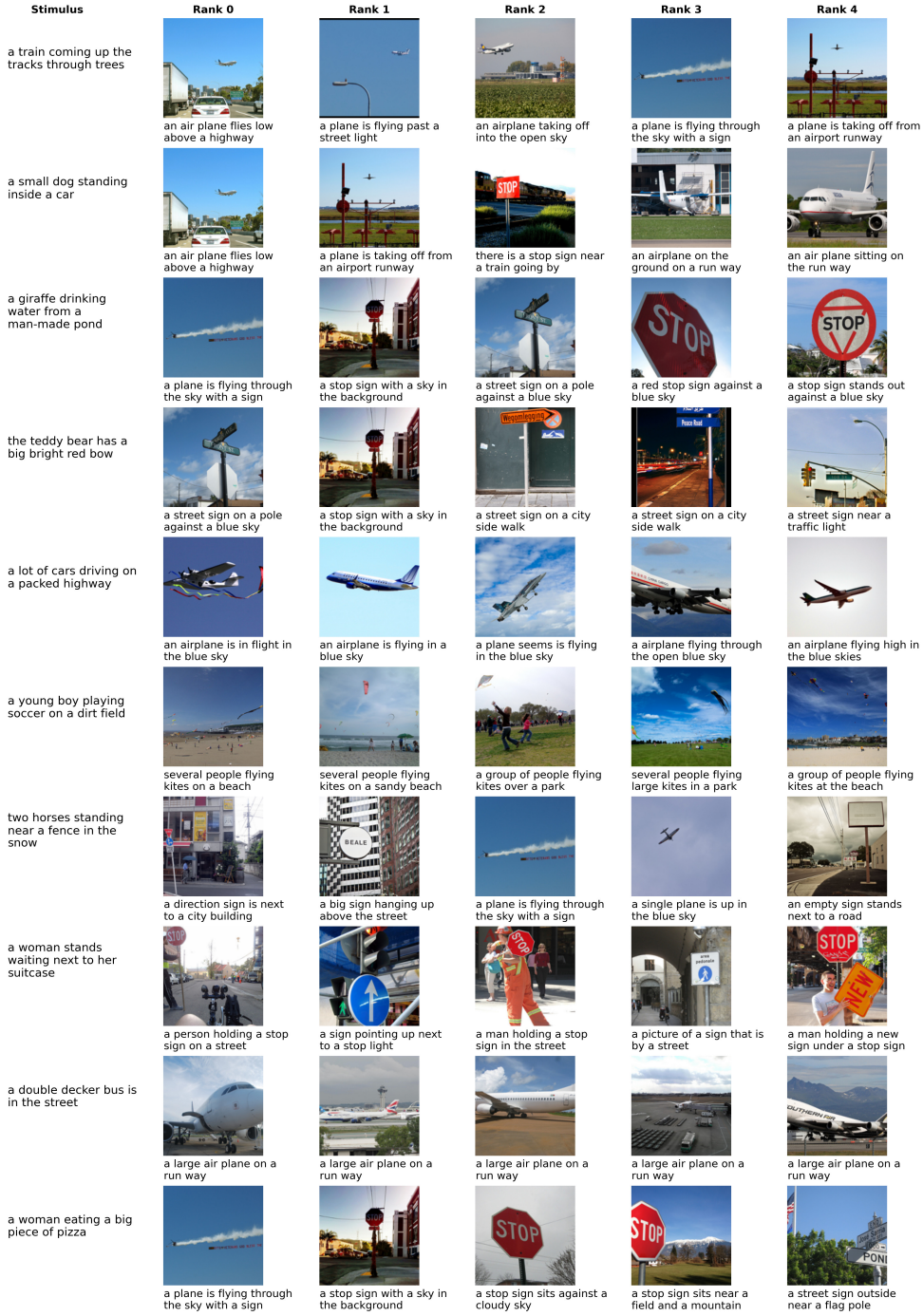
Appendix 3

Qualitative Decoding Results for Modality-Specific Decoders

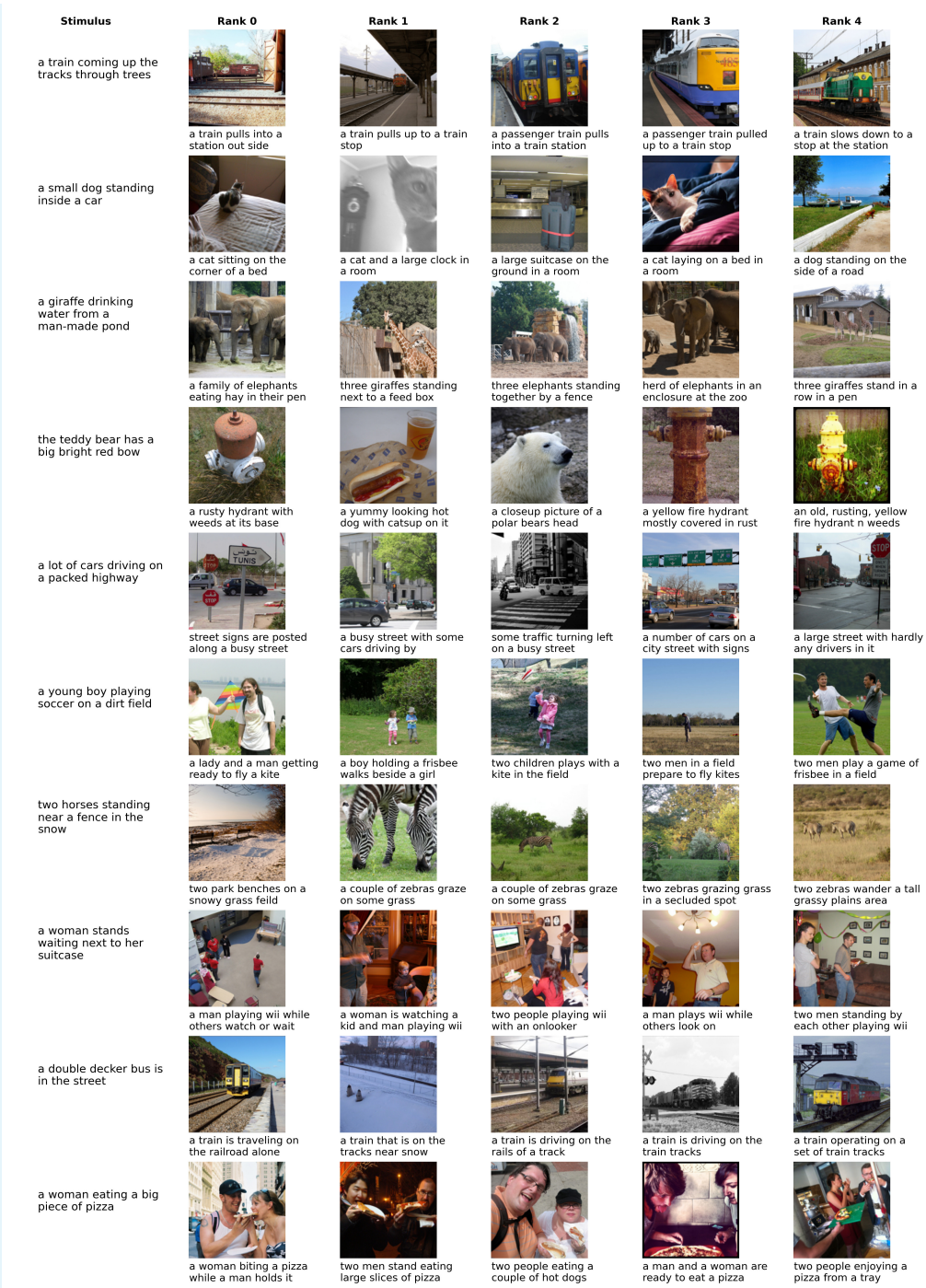
In this section we present qualitative decoding results for modality-specific decoders for the same examples as presented in Section Qualitative Decoding Results for modality-agnostic decoders. The results show that modality-agnostic decoders are as good as modality-specific decoders when evaluated in a within-modality decoding setup (Figures 1 and 3) but substantially better than modality-specific decoders when evaluated in a cross-decoding setup (Figures 2 and 4).



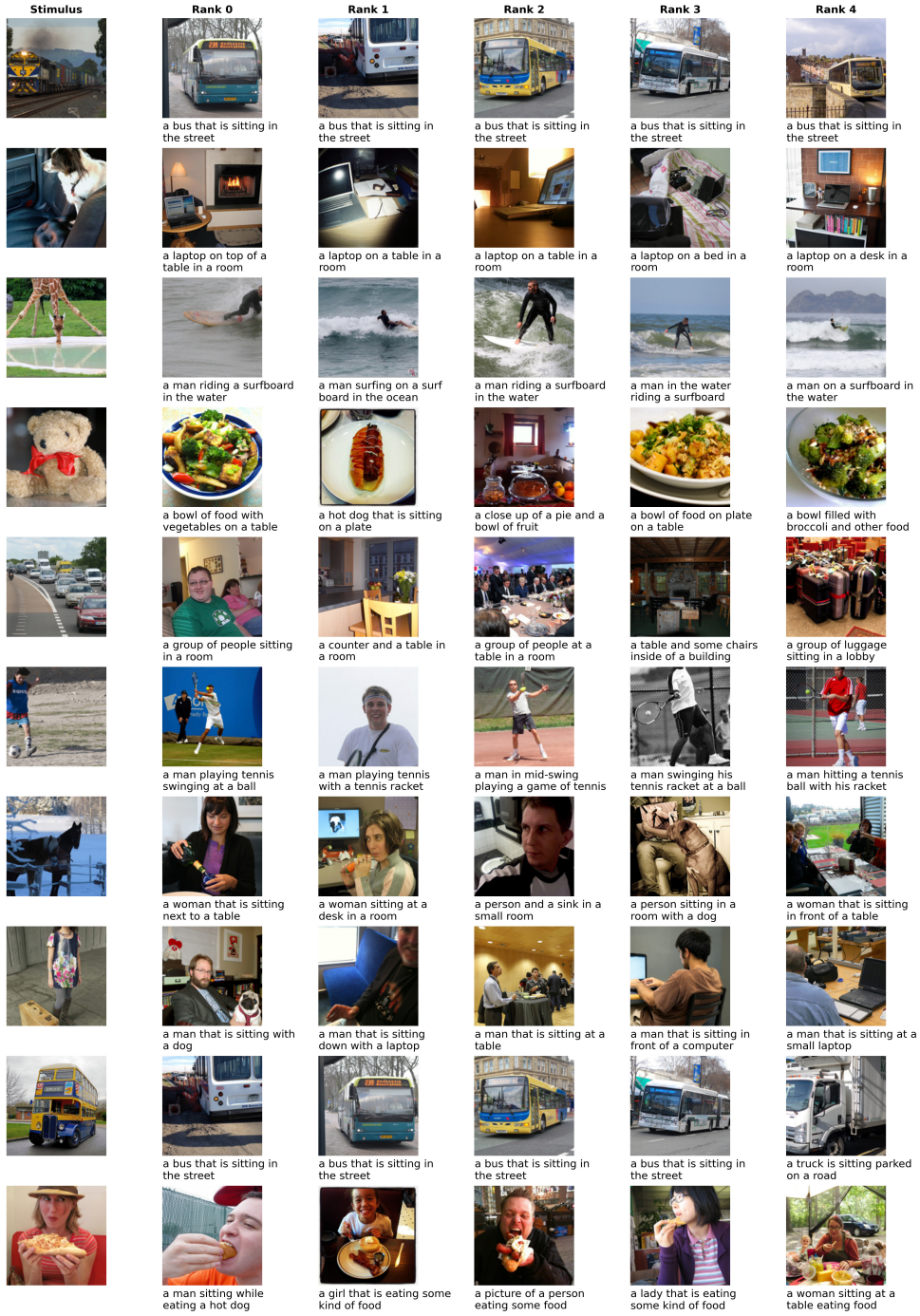
Appendix 3—figure 1. Decoding examples for image decoding using a modality-specific decoder trained on images (within-modality decoding). The first column shows the image the subject was seeing and the 5 following columns show the candidate stimuli with highest similarity to the predicted features, in descending order. We display both the image and the caption of the candidate stimuli because the decoder is based on multimodal features that are extracted from both modalities. All images were taken from the CoCo dataset (***Lin et al., 2014***).



Appendix 3—figure 2. Decoding examples for caption decoding using a modality-specific decoder trained on images (cross-modality decoding). For details see caption of Figure 1 All images were taken from the CoCo dataset (***Lin et al., 2014***).



Appendix 3—figure 3. Decoding examples for caption decoding using a modality-specific decoder trained on captions (within-modality decoding). For details see caption of Figure 1 All images were taken from the CoCo dataset (***Lin et al., 2014***).



Appendix 3—figure 4. Decoding examples for image decoding using a modality-specific decoder trained on captions (cross-modality decoding). For details see caption of Figure 1 All images were taken from the CoCo dataset (***Lin et al., 2014***).

Appendix 4

Candidates for Modality-invariant regions

Table 1 shows candidates for modality-invariant regions that were identified by previous work. We only considered fMRI experiments involving multiple stimulus modalities. Many other studies relied on unimodal stimuli and semantic tasks to identify modality-invariant regions, these are however less directly comparable to our setup.

Hemi	Region	Studies that identified the region as modality-invariant
left hemisphere	superior occipital gyrus*	(*Vandenberghe et al., 1996*; *Shinkareva et al., 2011*)
	middle occipital gyrus	(*Shinkareva et al., 2011*)
	inferior occipital gyrus	(*Shinkareva et al., 2011*; *Simanova et al., 2014*)
	superior temporal gyrus	(*Shinkareva et al., 2011*)
	superior temporal sulcus	(*Man et al., 2012*)
	middle temporal gyrus	(*Vandenberghe et al., 1996*; *Shinkareva et al., 2011*; *Fairhall and Caramazza, 2013*; *Devereux et al., 2013*; *Simanova et al., 2014*; *Handjaras et al., 2016*)
	inferior temporal gyrus	(*Vandenberghe et al., 1996*; *Shinkareva et al., 2011*; *Fairhall and Caramazza, 2013*; *Simanova et al., 2014*; *Handjaras et al., 2016*)
	fusiform gyrus	(*Vandenberghe et al., 1996*; *Moore and Price, 1999*; *Bright et al., 2004*; *Shinkareva et al., 2011*; *Fairhall and Caramazza, 2013*; *Simanova et al., 2014*)
	hippocampus	(*Vandenberghe et al., 1996*)
	parahippocampus	(*Bright et al., 2004*; *Fairhall and Caramazza, 2013*; *Handjaras et al., 2016*)
	perirhinal cortex	(*Bright et al., 2004*; *Fairhall and Caramazza, 2013*)
	superior parietal cortex	(*Shinkareva et al., 2011*)
	inferior parietal cortex	(*Shinkareva et al., 2011*; *Handjaras et al., 2016*)
	cingulate gyrus	(*Moore and Price, 1999*; *Fairhall and Caramazza, 2013*; *Handjaras et al., 2016*)
	cuneus	(*Shinkareva et al., 2011*)
	precuneus	(*Shinkareva et al., 2011*; *Handjaras et al., 2016*; *Fairhall and Caramazza, 2013*; *Popham et al., 2021*)
	supramarginal gyrus	(*Shinkareva et al., 2011*; *Handjaras et al., 2016*)
	angular gyrus	(*Shinkareva et al., 2011*; *Fairhall and Caramazza, 2013*; *Devereux et al., 2013*; *Simanova et al., 2014*; *Handjaras et al., 2016*; *Popham et al., 2021*)
	intraparietal sulcus	(*Shinkareva et al., 2011*; *Devereux et al., 2013*)
	temporoparietal junction	(*Vandenberghe et al., 1996*; *Handjaras et al., 2016*)
	superior frontal gyrus	(*Fairhall and Caramazza, 2013*)
	middle frontal gyrus	(*Fairhall and Caramazza, 2013*; *Handjaras et al., 2016*)
	inferior frontal gyrus	(*Vandenberghe et al., 1996*; *Bright et al., 2004*; *Simanova et al., 2014*; *Liuzzi et al., 2017*; *Handjaras et al., 2016*)
	precentral gyrus	(*Shinkareva et al., 2011*)
	postcentral gyrus	(*Shinkareva et al., 2011*)
	paracentral lobule	(*Shinkareva et al., 2011*)
	supplementary motor area	(*Shinkareva et al., 2011*)
right hemisphere	fusiform gyrus	(*Shinkareva et al., 2011*; *Simanova et al., 2014*)
	superior temporal sulcus	(*Man et al., 2012*)
	middle temporal gyrus	(*Handjaras et al., 2016*)
	inferior temporal gyrus	(*Handjaras et al., 2016*)
	angular gyrus	(*Handjaras et al., 2016*; *Popham et al., 2021*)
	superior parietal cortex	(*Shinkareva et al., 2011*)
	precuneus	(*Shinkareva et al., 2011*; *Popham et al., 2021*; *Fairhall and Caramazza, 2013*; *Handjaras et al., 2016*)
	cingulate gyrus	(*Moore and Price, 1999*; *Fairhall and Caramazza, 2013*; *Handjaras et al., 2016*; *Jung et al., 2018*)
	parahippocampus	(*Handjaras et al., 2016*)
	inferior parietal cortex	(*Handjaras et al., 2016*)
	paracentral lobule	(*Shinkareva et al., 2011*)
	middle frontal gyrus	(*Simanova et al., 2014*; *Jung et al., 2018*)
	superior frontal gyrus*	(*Jung et al., 2018*)
	inferior frontal gyrus	(*Moore and Price, 1999*; *Simanova et al., 2014*)

Appendix 4—table 1. Candidates for modality-invariant regions as identified by previous work. All these regions were also found in our analysis, except for the 2 regions marked with an asterisk (*).

Appendix 5

Searchlight size

In order to optimize the size of the searchlight for this analysis, we first ran a searchlight analysis with a fixed radius of 10mm using a modality-agnostic decoder on the data from the first subject (sub-01). Due to the shape of the cortex this leads of searchlights that contain varying numbers of vertices (on average: 897.4; max: 1580; min: 399). By observing the decoding scores as function of the number of vertices we find that performance peaks at 750 vertices (cf. Table 1). The final searchlight analyses was therefore performed with a searchlight of a fixed number of 750 vertices. The average radius with this number of vertices was 9.41mm (max: 13.65mm).

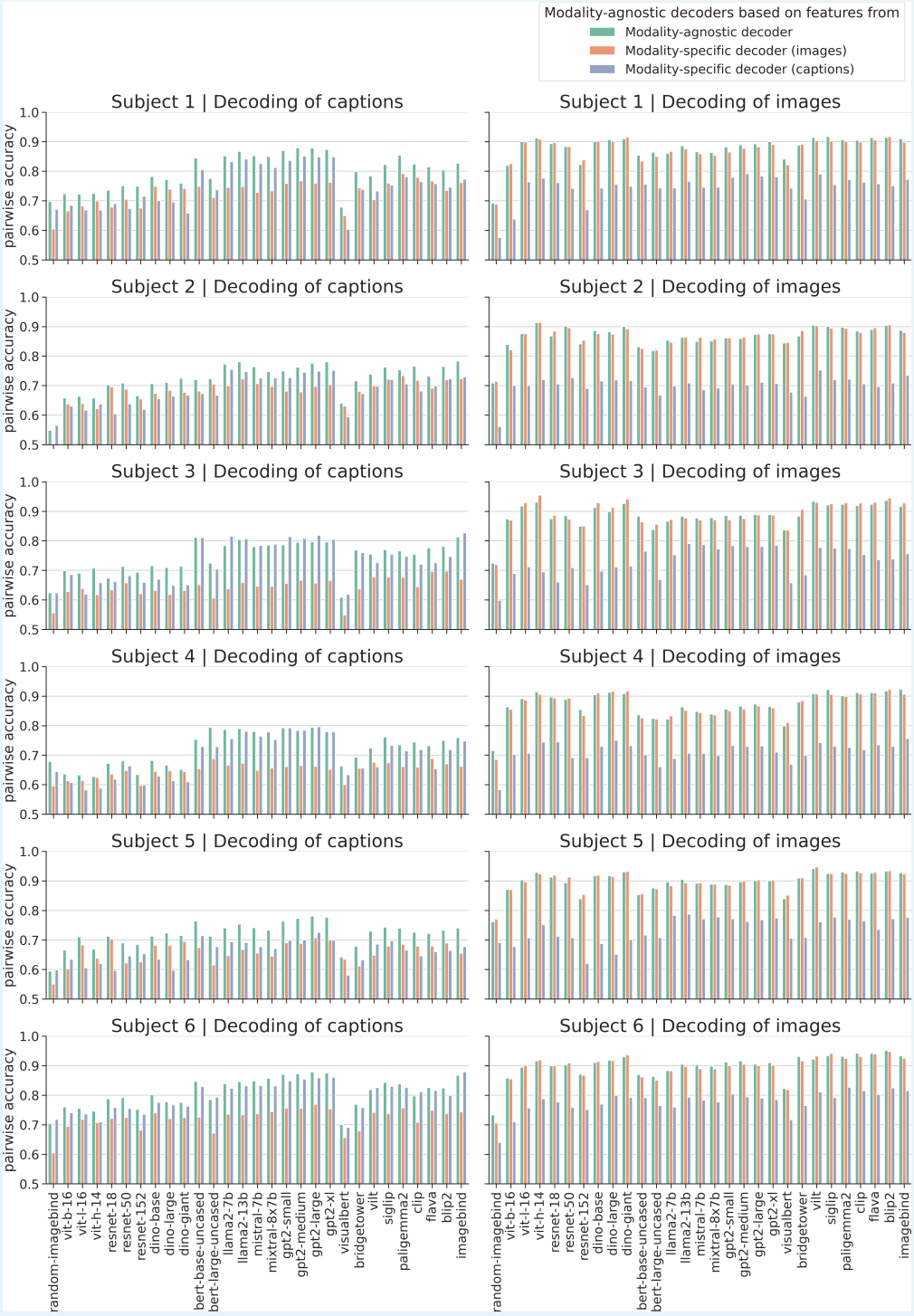
# vertices	pairwise accuracy (mean)
250	51.81%
500	52.35%
750	55.96%
1000	55.49%
1250	54.60%
1500	52.96%

Appendix 5—table 1. Average decoding scores (for images and captions) by number of vertices. Scores calculated based on the results of a searchlight analysis with a radius of 10mm. The accuracy values were grouped into bins based on the number of vertices that the searchlight was based on.

Appendix 6

Per-subject results

Results for individual subjects can be found in Figure 1. Among all subjects, we found similar converging results for decoding accuracies when comparing models, feature modalities, and modality-agnostic with modality-specific decoders.



Appendix 6—figure 1. Pairwise accuracy per subject. For details refer to Figure 5

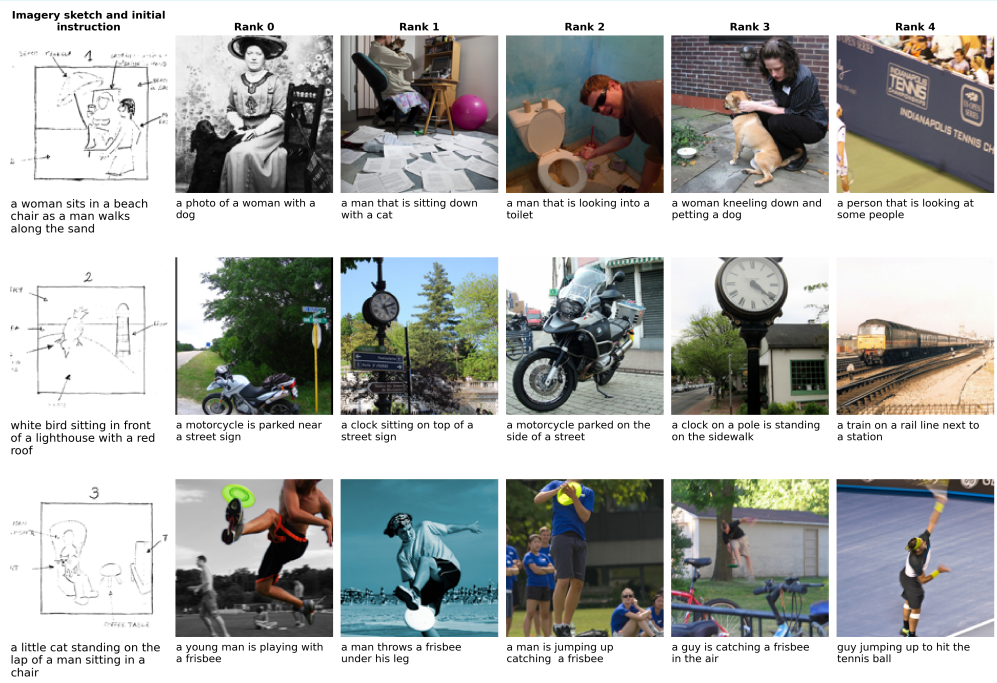
Appendix 7

Qualitative Imagery Decoding Results

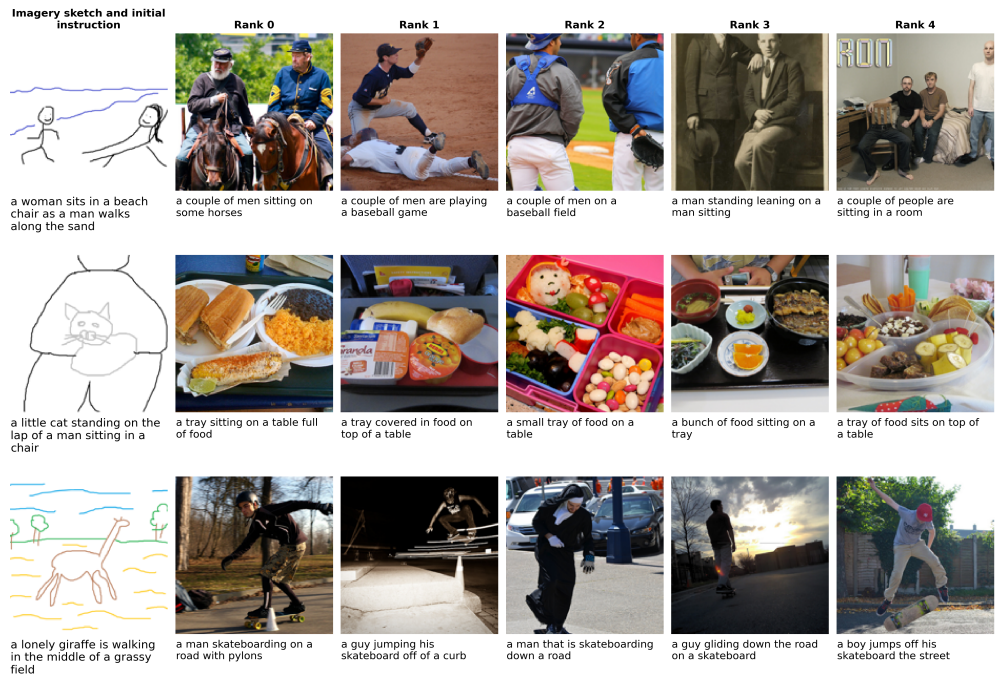
The following figures 1-6 present qualitative decoding results for the imagery conditions using a modality-agnostic decoder. We present the results separately for each subject as each subject chose an individual set of 3 stimuli to perform mental imagery on. In each plot, the leftmost column shows the caption that was used as initial instruction as well as the subject's sketch that they drew at the end of the experiment.

We used the same large candidate set of 41K stimuli as for the qualitative decoding results for the other conditions (see also Section Qualitative Decoding Results).

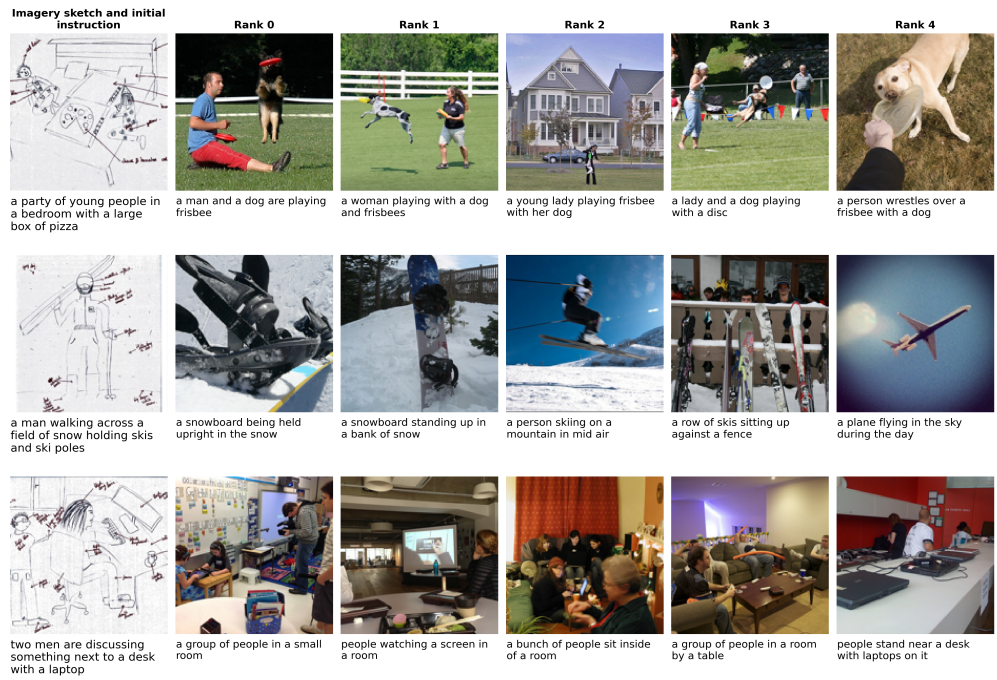
Overall, we found that the decoding quality for imagery stimuli lags behind that for trials with perceived stimuli. This was expected and confirms the quantitative results reported in Section Imagery Decoding. Still, in several cases some of the concepts are decoded correctly (e.g. a women in the first row of Figure 1; winter sports in the second row of Figure 3; laptops/screens and multiple people in the third row of Figure 3).



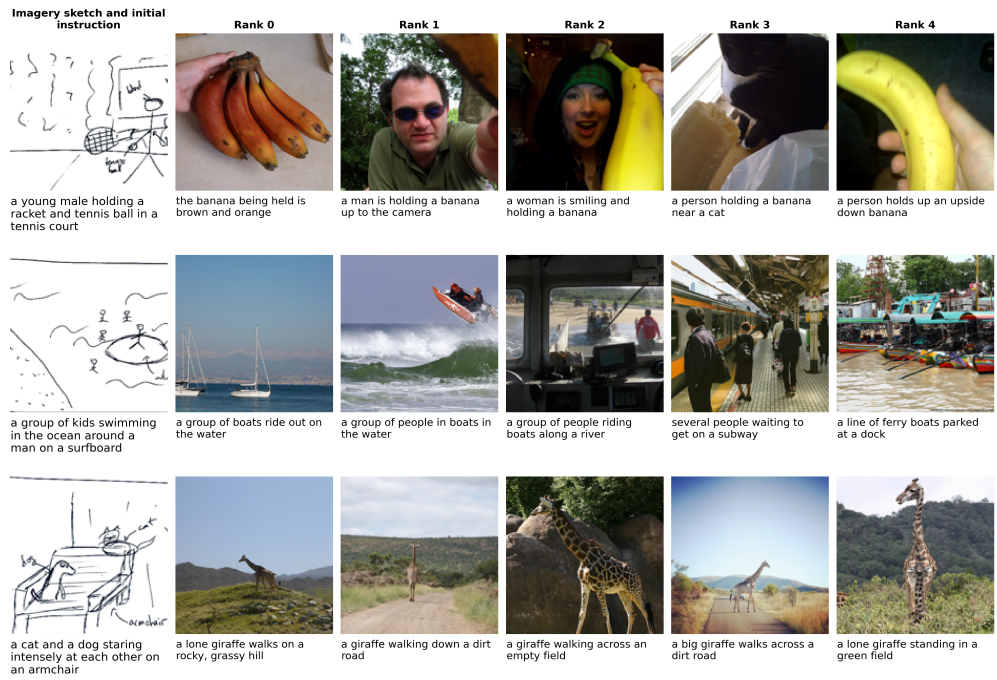
Appendix 7—figure 1. Imagery decoding for Subject 1 using a modality-agnostic decoder. The first column shows the caption that was used to stimulate imagery and on top of it the sketch of their mental image that the subjects draw at the end of the experiment. The 5 following columns show the candidate stimuli with highest similarity to the predicted features, in descending order. We display both the image and the caption of the candidate stimuli because the decoder is based on multimodal features that are extracted from both modalities. All images were taken from the CoCo dataset (***Lin et al., 2014***).



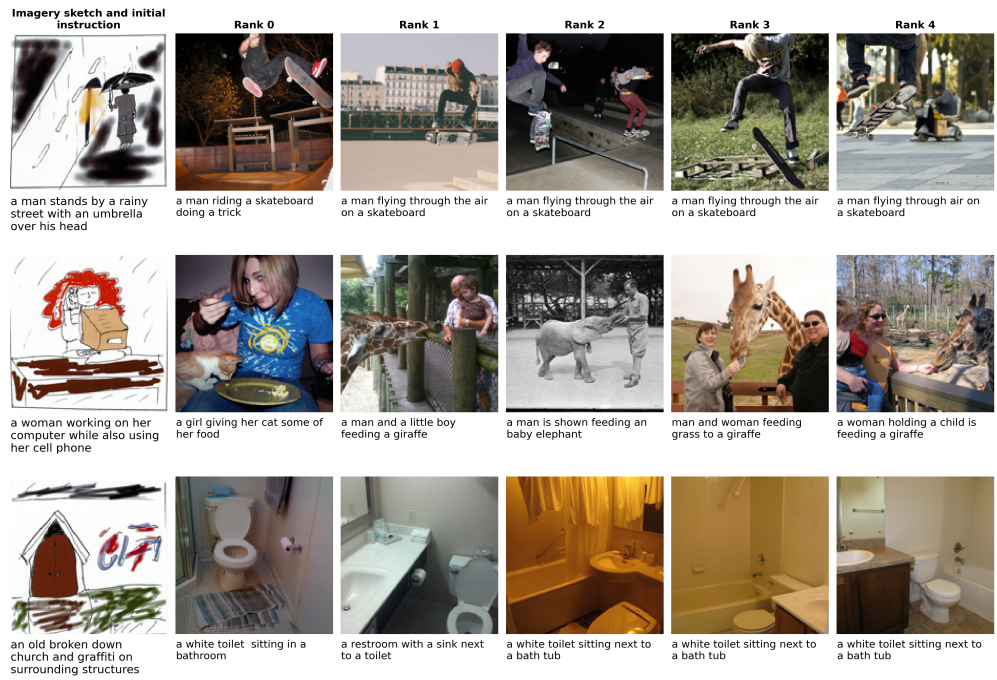
Appendix 7—figure 2. Imagery decoding for Subject 2 using a modality-agnostic decoder. For further details refer to the caption of Figure 1. All images were taken from the CoCo dataset (***Lin et al., 2014***).



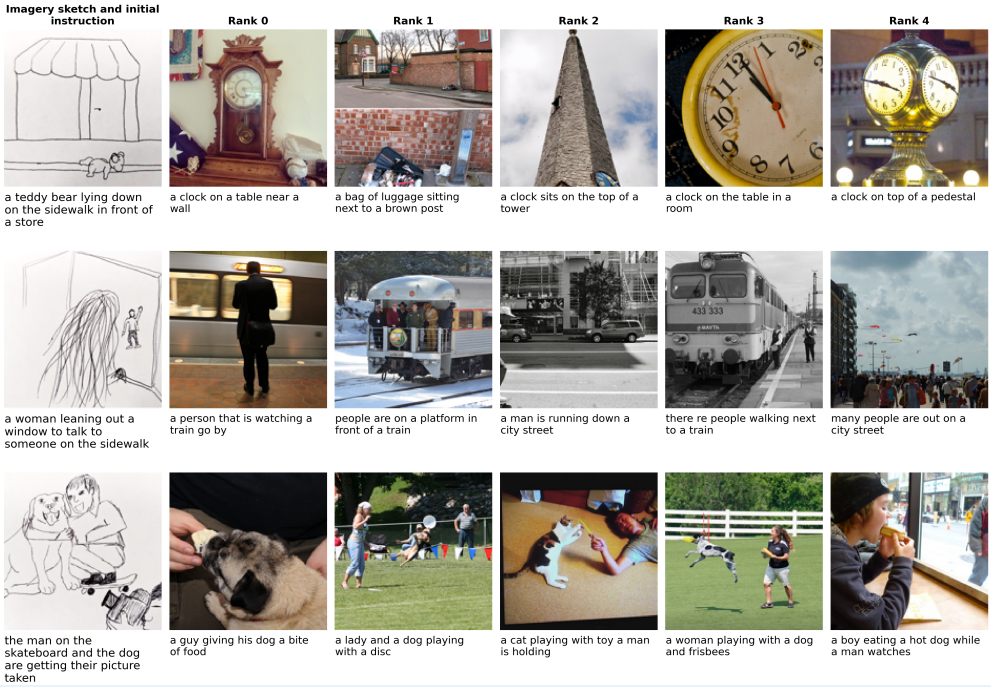
Appendix 7—figure 3. Imagery decoding for Subject 3 using a modality-agnostic decoder. For further details refer to the caption of Figure 1. All images were taken from the CoCo dataset (***Lin et al., 2014***).



Appendix 7—figure 4. Imagery decoding for Subject 4 using a modality-agnostic decoder. For further details refer to the caption of Figure 1. All images were taken from the CoCo dataset (***Lin et al., 2014***).



Appendix 7—figure 5. Imagery decoding for Subject 5 using a modality-agnostic decoder. For further details refer to the caption of Figure 1. All images were taken from the CoCo dataset (***Lin et al., 2014***).



Appendix 7—figure 6. Imagery decoding for Subject 6 using a modality-agnostic decoder. For further details refer to the caption of Figure 1. All images were taken from the CoCo dataset (***Lin et al., 2014***).