BRAIN DECODING: TOWARD REAL-TIME RECONSTRUCTION OF VISUAL PERCEPTION

Yohann Benchetrit[1,*], **Hubert Banville**[1,*], **Jean-Rémi King**[1,2]
[1]FAIR, Meta, [2] Laboratoire des Systèmes Perceptifs, École Normale Supérieure, PSL University
{ybenchetrit,hubertjb,jeanremi}@meta.com

ABSTRACT

In the past five years, the use of generative and foundational AI systems has greatly improved the decoding of brain activity. Visual perception, in particular, can now be decoded from functional Magnetic Resonance Imaging (fMRI) with remarkable fidelity. This neuroimaging technique, however, suffers from a limited temporal resolution (≈0.5 Hz) and thus fundamentally constrains its real-time usage. Here, we propose an alternative approach based on magnetoencephalography (MEG), a neuroimaging device capable of measuring brain activity with high temporal resolution (≈5,000 Hz). For this, we develop an MEG decoding model trained with both contrastive and regression objectives and consisting of three modules: i) pretrained embeddings obtained from the image, ii) an MEG module trained end-to-end and iii) a pretrained image generator. Our results are threefold: Firstly, our MEG decoder shows a 7X improvement of image-retrieval over classic linear decoders. Second, late brain responses to images are best decoded with DINOv2, a recent foundational image model. Third, image retrievals and generations both suggest that MEG signals primarily contain high-level visual features, whereas the same approach applied to 7T fMRI also recovers low-level features. Overall, these results provide an important step towards the decoding – in real time – of the visual processes continuously unfolding within the human brain.

1 INTRODUCTION

Automating the discovery of brain representations. Understanding how the human brain represents the world is arguably one of the most profound scientific challenges. This quest, which originally consisted of searching, one by one, for the specific features that trigger each neuron, (*e.g.* Hubel & Wiesel (1962); O'Keefe & Nadel (1979); Kanwisher et al. (1997)), is now being automated by Machine Learning (ML) in two mains ways. First, as a signal processing *tool*, ML algorithms are trained to extract informative patterns of brain activity in a data-driven manner. For example, Kamitani & Tong (2005) trained a support vector machine to classify the orientations of visual gratings from functional Magnetic Resonance Imaging (fMRI). Since then, deep learning has been increasingly used to discover such brain activity patterns (Roy et al., 2019; Thomas et al., 2022; Jayaram & Barachant, 2018; Défossez et al., 2022; Scotti et al., 2023). Second, ML algorithms are used as functional *models* of the brain. For example, Yamins et al. (2014) have shown that the embedding of natural images in pretrained deep nets linearly account for the neuronal responses to these images in the cortex. Since, pretrained deep learning models have been shown to account for a wide variety of stimuli including text, speech, navigation, and motor movement (Banino et al., 2018; Schrimpf et al., 2020; Hausmann et al., 2021; Mehrer et al., 2021; Caucheteux et al., 2023).

Generating images from brain activity. This observed representational alignment between brain activity and deep learning models creates a new opportunity: Decoding of visual stimuli need not be restricted to a limited set of classes, but can now leverage pretrained representations to condition subsequent generative AI models. While the resulting image may be partly "hallucinated", interpreting images can be much simpler than interpreting latent features. Following a long series

[*]Equal contribution.

of generative approaches (Nishimoto et al., 2011; Kamitani & Tong, 2005; VanRullen & Reddy, 2019; Seeliger et al., 2018), diffusion techniques have, in this regard, significantly improved the generation of images from functional Magnetic Resonance Imaging (fMRI). The resulting pipeline typically consists of three main modules: (1) a set of pretrained embeddings obtained from the image onto which (2) fMRI activity can be linearly mapped and (3) ultimately used to condition a pretrained image-generation model (Ozcelik & VanRullen, 2023; Mai & Zhang, 2023; Zeng et al., 2023; Ferrante et al., 2022). These recent fMRI studies primarily differ in the type of pretrained image-generation model that they use.

The challenge of real-time decoding. This generative decoding approach has been mainly applied to fMRI. However, the temporal resolution of fMRI is limited by the time scale of blood flow and typically leads to one snapshot of brain activity every two seconds – a time scale that challenges its clinical usage, *e.g.* for patients who require a brain-computer-interface (Willett et al., 2023; Moses et al., 2021; Metzger et al., 2023; Défossez et al., 2022). On the contrary, magnetoencephalography (MEG) can measure brain activity at a much higher temporal resolution (≈5,000 Hz) by recording the fluctuation of magnetic fields elicited by the post-synaptic potentials of pyramidal neurons. This higher temporal resolution comes at cost, however: the spatial resolution of MEG is limited to ≈300 sensors, whereas fMRI measures ≈100,000 voxels. In sum, fMRI intrinsically limits our ability to (1) track the dynamics of neuronal activity, (2) decode dynamic stimuli (speech, videos etc) and (3) apply these tools to real-time use cases. Conversely, it is unknown whether temporally-resolved neuroimaging systems like MEG are sufficiently precise to generate natural images in real-time.

Our approach. Combining previous work on speech retrieval from MEG (Défossez et al., 2022) and on image generation from fMRI (Takagi & Nishimoto, 2023; Ozcelik & VanRullen, 2023), we here develop a three-module pipeline trained to (1) align MEG activity onto pretrained visual embeddings and (2) generate images from a stream of MEG signals (Fig. 1).

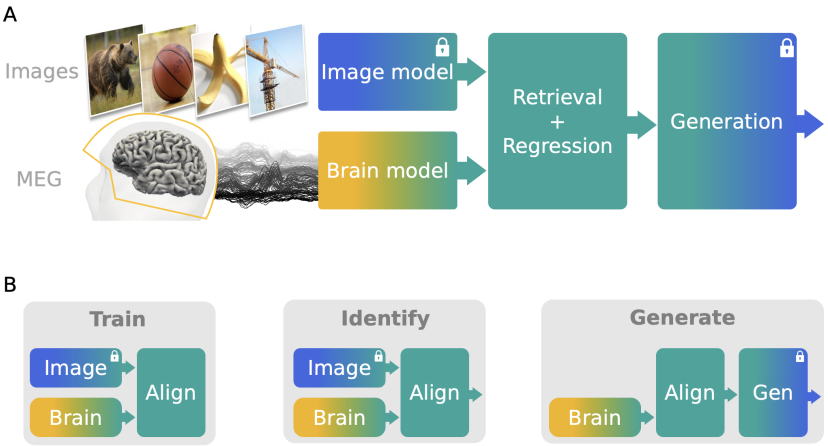

Figure 1: (**A**) Approach. Locks indicate pretrained models. (**B**) Processing schemes. Unlike image generation, image retrieval can be done in the aligned latent space, but requires the true image in the retrieval set.

Our systematic benchmark provides two main contributions: our MEG decoder leads to (1) high-performing image retrieval and image generation, (2) new means to interpret the unfolding of visual processing in the brain. This demonstrates the capacity of our approach to truly generalize to new visual concepts, paving the way to "free-form" visual decoding. Overall, our findings outline a promising avenue for real-time decoding of visual representations in the lab and in the clinic.

2 METHODS

2.1 PROBLEM STATEMENT

We aim to decode images from multivariate time series of brain activity recorded with MEG as healthy participants watched a sequence of natural images. Let $\boldsymbol{X}_i \in \mathbb{R}^{C \times T}$ be the MEG time window collected as an image I_i was presented to the participant, where C is the number of MEG channels, T is the number of time points in the MEG window and $i \in [\![1, N]\!]$. Let $\boldsymbol{z}_i \in \mathbb{R}^F$ be the latent representation of I_i, with F the number of features, obtained by embedding the image using a pretrained image model (Section 2.4). As described in more detail below, our decoding approach relies on training a *brain module* $\mathbf{f}_\theta : \mathbb{R}^{C \times T} \to \mathbb{R}^F$ to maximally retrieve or predict I_i through $\boldsymbol{z}_i$, given $\boldsymbol{X}_i$.

2.2 TRAINING OBJECTIVES

We use different training objectives for the different parts of our proposed pipeline. First, in the case of retrieval, we aim to pick the right image I_i (*i.e.*, the one corresponding to $\boldsymbol{X}_i$) out of a bank of candidate images. To do so, we train $\mathbf{f}_\theta$ using the CLIP loss (Radford et al., 2021) on batches of size B with exactly one positive example:

$$\mathcal{L}_{CLIP}(\theta) = -\frac{1}{B}\sum_{i=1}^{B}\left(\log\frac{\exp(s(\hat{\boldsymbol{z}}_i, \boldsymbol{z}_i)/\tau)}{\sum_{j=1}^{B}\exp(s(\hat{\boldsymbol{z}}_i, \boldsymbol{z}_j)/\tau)} + \log\frac{\exp(s(\hat{\boldsymbol{z}}_i, \boldsymbol{z}_i)/\tau)}{\sum_{k=1}^{B}\exp(s(\hat{\boldsymbol{z}}_k, \boldsymbol{z}_i)/\tau)}\right) \tag{1}$$

where s is the cosine similarity, $\boldsymbol{z}_i$ and $\hat{\boldsymbol{z}}_i = \mathbf{f}_\theta(\boldsymbol{X}_i)$ are the latent representation and the corresponding MEG-based prediction, respectively, and τ is a learned temperature parameter.

Next, to go beyond retrieval and instead generate images, we train $\mathbf{f}_\theta$ to directly predict the latent representations $\boldsymbol{z}$ such that we can use them to condition generative image models. This is done using a standard mean squared error (MSE) loss:

$$\mathcal{L}_{MSE}(\theta) = \frac{1}{NF}\sum_{i=1}^{N}\|\boldsymbol{z}_i - \hat{\boldsymbol{z}}_i\|_2^2 \tag{2}$$

Finally, we combine the CLIP and MSE losses using a convex combination with tuned weight to train models that benefit from both training objectives:

$$\mathcal{L}_{Combined} = \lambda\mathcal{L}_{CLIP} + (1-\lambda)\mathcal{L}_{MSE} \tag{3}$$

2.3 BRAIN MODULE

We adapt the dilated residual ConvNet architecture of Défossez et al. (2022), denoted as $\mathbf{f}_\theta$, to learn the projection from an MEG window $\boldsymbol{X}_i \in \mathbb{R}^{C \times T}$ to a latent image representation $\boldsymbol{z}_i \in \mathbb{R}^F$. The original model's output $\hat{\boldsymbol{Y}}_{backbone} \in \mathbb{R}^{F' \times T}$ maintains the temporal dimension of the network through its residual blocks. However, here we regress a single latent per input instead of a sequence of T latents like in Défossez et al. (2022). Consequently, we add a temporal aggregation layer to reduce the temporal dimension of $\hat{\boldsymbol{Y}}_{backbone}$ to obtain $\hat{\boldsymbol{y}}_{agg} \in \mathbb{R}^{F'}$. We experiment with three types of aggregations: global average pooling, a learned affine projection, and an attention layer. Finally, we add two MLP heads[1], *i.e.*, one for each term in $\mathcal{L}_{Combined}$, to project from F' to the F dimensions of the target latent.

We run a hyperparameter search to identify an appropriate configuration of preprocessing, brain module architecture, optimizer and loss hyperparameters for the retrieval task (see Appendix A.2). The final architecture configuration for retrieval is described in Table S2 and contains *e.g.* 6.4M trainable parameters for $F = 768$.

[1] A head consists of repeated LayerNorm-GELU-Linear blocks.

For image generation experiments, the output of the MSE head is further postprocessed as in Ozcelik & VanRullen (2023), *i.e.*, we z-score normalize each feature across predictions, and then apply the inverse z-score transform fitted on the training set (defined by the mean and standard deviation of each feature dimension on the target embeddings). We select λ in $\mathcal{L}_{Combined}$ by sweeping over $\{0.0, 0.25, 0.5, 0.75, 1.0\}$ and pick the model whose top-5 accuracy is the highest on the large test set. Of note, when training models to generate CLIP and AutoKL latents, we simplify the task of the CLIP head by reducing the dimensionality of its target: we use the CLS token for CLIP-Vision ($F_{MSE} = 768$), the "mean" token for CLIP-Text ($F_{MSE} = 768$), and the channel-average for AutoKL latents ($F_{MSE} = 4096$), respectively.

2.4 IMAGE MODULES

We study the functional alignment between brain activity and a variety of (output) embeddings obtained from deep neural networks trained in three different representation learning paradigms, spanning a wide range of dimensionalities: supervised learning (*e.g.* VGG-19), image-text alignment (CLIP), and variational autoencoders. When using vision transformers, we further include two additional embeddings of smaller dimensionality: the average of all output embeddings across tokens (mean), and the output embedding of the class-token (CLS). For comparison, we also evaluate our approach on human-engineered features obtained without deep learning. The list of embeddings is provided in Appendix A.4. For clarity, we focus our experiments on a representative subset.

2.5 GENERATION MODULE

To fairly compare our work to the results obtained with fMRI results, we follow the approach of Ozcelik & VanRullen (2023) and use a model trained to generate images from pretrained embeddings. Specifically, we use a latent diffusion model conditioned on three embeddings: CLIP-Vision (257 tokens $\times$ 768), CLIP-Text (77 tokens $\times$ 768), and a variational autoencoder latent (AutoKL; $(4 \times 64 \times 64)$. Following Ozcelik & VanRullen (2023), we apply diffusion with 50 DDIM steps, a guidance of 7.5, a strength of 0.75 with respect to the image-to-image pipeline, and a mixing of 0.4.

2.6 TRAINING AND COMPUTATIONAL CONSIDERATIONS

Cross-participant models are trained on a set of $\approx 63,000$ examples using the Adam optimizer (Kingma & Ba, 2014) with learning rate of 3×10^{-4} and a batch size of 128. We use early stopping on a validation set of $\approx 15,800$ examples randomly sampled from the original training set, with a patience of 10, and evaluate the performance of the model on a held-out test set (see below). Models are trained on a single Volta GPU with 32 GB of memory. We train each model three times using three different random seeds for the weight initialization of the brain module.

2.7 EVALUATION

Retrieval metrics. We first evaluate decoding performance using retrieval metrics. For a known test set, we are interested in the probability of identifying the correct image given the model predictions. Retrieval metrics have the advantage of sharing the same scale regardless of the dimensionality of the MEG (like encoding metrics), the dimensionality of the image embedding (like regression metrics). We evaluate retrieval using either the *relative median rank* (which does not depend on the size of the retrieval set), defined as the rank of a prediction divided by the size of the retrieval set, or the *top-5 accuracy* (which is more common in the literature).

Generation metrics. Decoding performance is often measured qualitatively as well as quantitatively using a variety of metrics reflecting the reconstruction fidelity both in terms of perception and semantics. For fair comparison with fMRI generations, we provide the same metrics as Ozcelik & VanRullen (2023), computed between seen and generated images: PixCorr (the pixel-wise correlation between the true and generated images), SSIM (Structural Similarity Index Metric), and SwAV (the correlation with respect to SwAV-ResNet50 output). On the other hand, AlexNet(2/5), Inception, and CLIP are the respective 2-way comparison scores of layers 2/5 of AlexNet, the pooled last layer of Inception and the output layer of CLIP. For the NSD dataset, these metrics are reported for participant 1 only (see Appendix A.5).

To avoid non-representative cherry-picking, we sort all generations on the test set according to the sum of (minus) SwAV and SSIM. We then split the data into 15 blocks and pick 4 images from the best, middle and worst blocks with respect to the summed metric.

Real-time and average metrics. It is common in fMRI to decode brain activity from preprocessed values estimated with a General Linear Model. These "beta values" are estimates of brain responses to individual images, computed across multiple repetitions of such images. To provide a fair assessment of possible MEG decoding performance, we thus leverage repeated image presentations available in the datasets (see below) by averaging predictions before evaluating metrics.

2.8 DATASET

We test our approach on the "THINGS-MEG" dataset (Hebart et al., 2023). Four participants (2 females, 2 males; mean age of 23.25 years), underwent 12 MEG sessions during which they were presented with a set of 22,448 unique images selected from the THINGS database (Hebart et al., 2019), covering 1,854 categories. Of those, only a subset of 200 images (each one of a different category) was shown multiple times to the participants. The images were displayed for 500 ms each, with a variable fixation period of 1000 ± 200 ms between presentations. The THINGS dataset additionally contains 3,659 images that were not shown to the participants and that we use to augment the size of our retrieval set and emphasize the robustness of our method.

MEG Preprocessing. We use a minimal MEG data-preprocessing pipeline as in Défossez et al. (2022). Raw data from the 272 MEG radial gradiometer channels is downsampled from 1,200 Hz to 120 Hz before being centered and clipped channel-wise above +/- 5 standard errors. The continuous MEG data is then epoched from -500 ms to 1,000 ms relative to stimulus onset. Finally, baseline-correction is performed by subtracting the mean signal value observed between the start of an epoch and the stimulus onset for each channel.

Splits. The original split of Hebart et al. (2023) consists of 22,248 uniquely presented images, and 200 test images repeated 12 times each for each participant (*i.e.*, 2,400 trials per participant). The use of this data split presents a challenge, however, as the test set contains only one image per category, and these categories are also seen in the training set. This means evaluating retrieval performance on this test set does not measure the capacity of the model to (1) extrapolate to new unseen categories of images and (2) recover a particular image within a set of multiple images of the same category, but rather only to "categorize" it. Consequently, we propose two modifications of the original split. First, we remove from the training set any image whose category appears in the original test set. This "adapted training set" removes any categorical leakage across the train/test split and makes it possible to assess the capacity of the model to decode images of unseen image categories (*i.e.*, a "zero-shot" setting). Second, we propose a new "large test set" that is built using the images removed from the training set. This new test set effectively allows evaluating retrieval performance of images within images of the same category[2]. We report results on both the original ("small") and the "large" test sets to enable comparisons with the original settings of Ozcelik & VanRullen (2023). Finally, we also compare our results to the performance obtained by a similar pipeline but trained on fMRI data using the NSD dataset (Allen et al., 2022) (see Appendix A.5).

3 RESULTS

ML as an effective *model* of the brain. Which representations of natural images are likely to maximize decoding performance? To answer this question, we compare the retrieval performance obtained by linear Ridge regression models trained to predict one of 16 different latent visual representations given the flattened MEG response $\boldsymbol{X}_i$ to each image I_i (Table S1). While all image embeddings lead to above-chance retrieval, supervised and text/image alignment models (*e.g.* VGG, CLIP) yield the highest retrieval scores.

[2]We leave out images of the original test set from this new large test set, as keeping them would create a discrepancy between the number of MEG repetitions for training images and test images.

ML as an effective *tool* to learn brain responses. We then compare these linear baselines to a deep ConvNet architecture (Défossez et al., 2022) trained on the same task[3], *i.e.*, to retrieve the matching image given an MEG window. Using a deep model leads to a 7X improvement over the linear baselines (Fig. 2). Multiple types of image embeddings lead to good retrieval performance, with VGG-19 (supervised learning), CLIP-Vision (text/image alignment) and DINOv2 (self-supervised learning) yielding top-5 accuracies of $70.33 \pm 2.80\%$, $68.66 \pm 2.84\%$, $68.00 \pm 2.86\%$, respectively (where the standard error of the mean is computed across the averaged image-wise metrics). Similar conclusions, although with lower performance, can be drawn from our "large" test set setting, where decoding cannot rely solely on the image category but also requires discriminating between multiple images of the same category. Representative retrieval examples are shown in Appendix A.3.

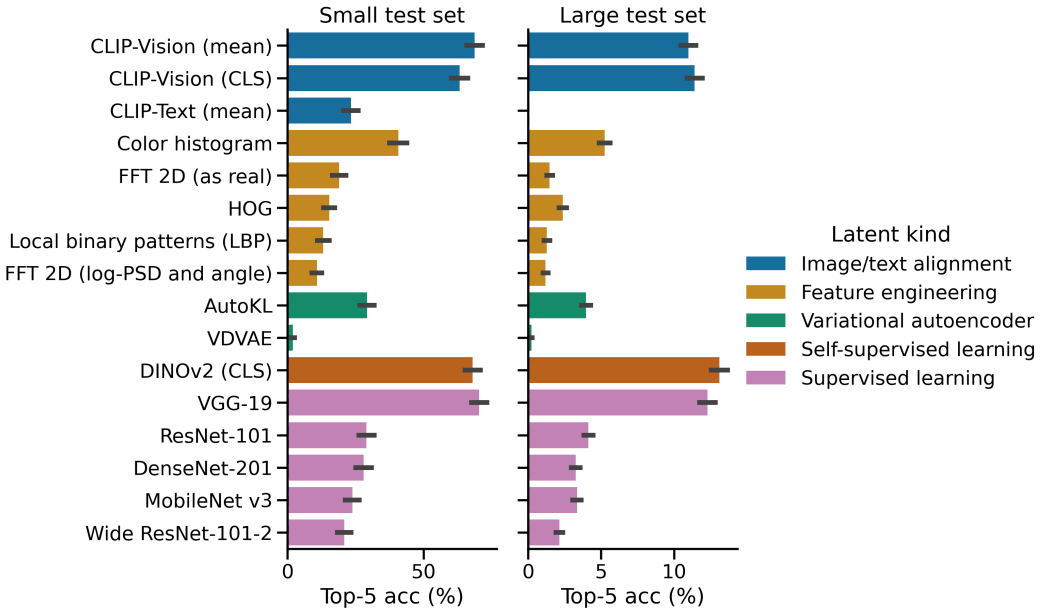

Figure 2: Image retrieval performance obtained from a trained deep ConvNet. The original "small" test set (Hebart et al., 2023) comprises 200 distinct images, each belonging to a different category. In contrast, our proposed "large" test set comprises 12 images from each of those 200 categories, yielding a total of 2,400 images. Chance-level is 2.5% top-5 accuracy for the small test set and 0.21% for the large test set. The best latent representations yield accuracies around 70% and 13% for the small and large test sets, respectively.

Temporally-resolved image retrieval. The above results are obtained from the full time window (-500 ms to 1,000 ms relative to stimulus onset). To further investigate the possibility of decoding visual representations as they unfold in the brain, we repeat this analysis on 250 ms-long sliding windows (Fig. 3). For clarity, we focus on a subset of representative image embeddings. As expected, all models yield chance-level performance before the image presentation. For all models, a first clear peak can then be observed on the 0 to 250-ms window, followed by a second peak, after the image offset, which then quickly goes back to chance-level. Interestingly, the recent self-supervised model DINOv2 yields particularly good retrieval performance after the image offset.

To get a better sense of what the above decoding metrics mean, we present the top-1 retrieved images from an augmented retrieval set built by concatenating the "large" test set with an additional set of 3,659 images that were not seen by the participants (Fig. 4).

Overall, the retrieved images tend to come from the correct category, such as "speaker" or "brocoli", mostly during the first few sub-windows ($t \leq 1$ s). However, these retrieved images do not appear to share obvious low-level features to the images seen by the participants.

[3]We use $\lambda = 1$ in $\mathcal{L}_{Combined}$ as we are solely concerned with the retrieval part of the pipeline here.

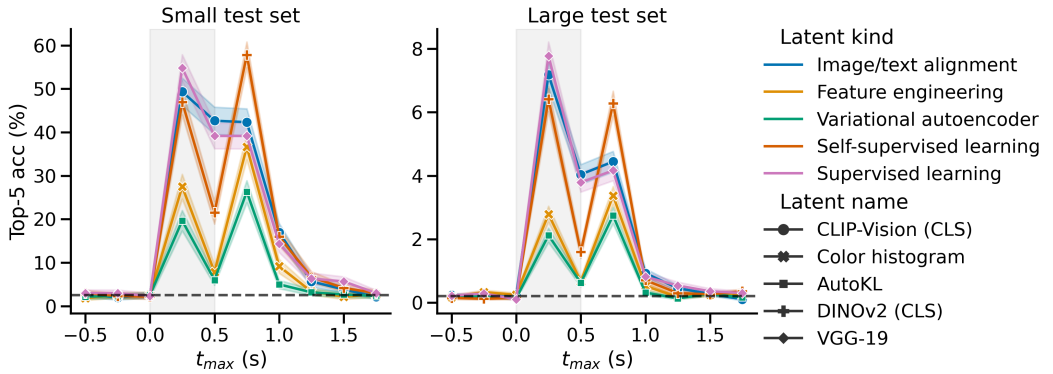

Figure 3: Retrieval performance of models trained on 250 ms sliding windows for different image embeddings. The shaded gray area indicates the 0.5-s interval during which the image was presented to the participants. Accuracy generally peaked right after the image onset and offset.

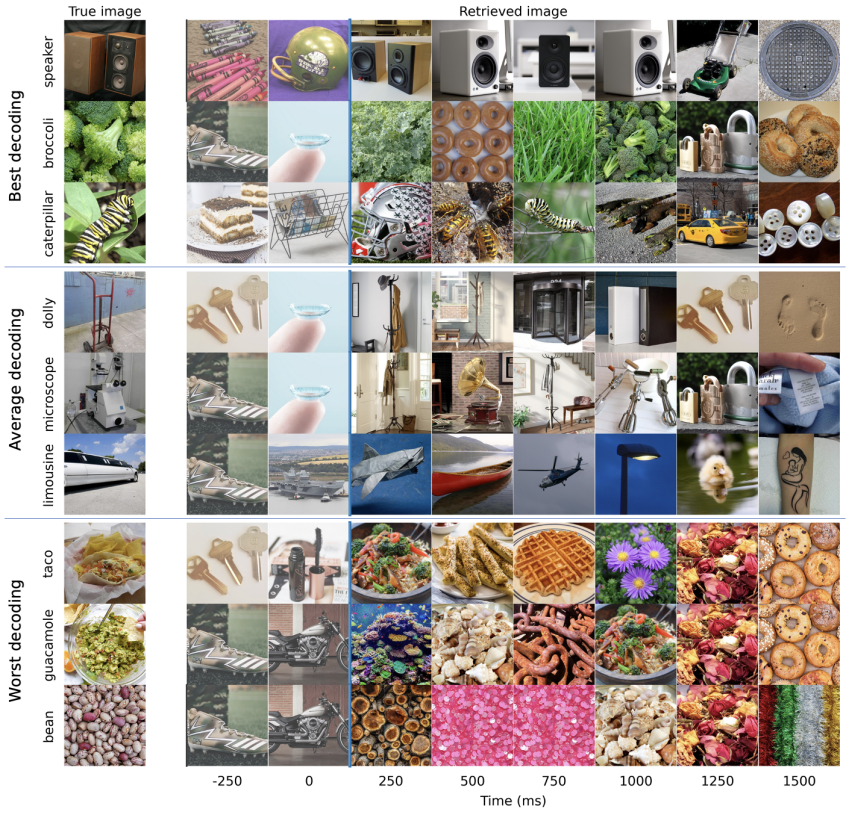

Figure 4: Representative examples of dynamic retrievals using CLIP-Vision (CLS) and models trained on 250-ms sliding windows (Image onset: $t = 0$, retrieval set: $N = 6,059$ from $1,196$ categories). The groups of three stacked rows represent best, average and worst retrievals, obtained by sampling examples from the $<$10%, 45-55% and $>$90% percentile groups based on top-5 accuracy.

Overall, and while further analyses of these results remain necessary, it seems that (1) our decoding leverages the brain responses related to both the onset and the offset of the image and (2) category-level information dominates these visual representations as early as 250 ms.

Table 1: Quantitative evaluation of reconstruction quality from MEG data on THINGS-MEG (compared to fMRI data on NSD (Allen et al., 2022) using a cross-validated Ridge regression). We report PixCorr, SSIM, AlexNet(2), AlexNet(5), Inception, SwAV and CLIP (the side-arrow indicates whether better scores are higher or lower). In particular, this shows that fMRI betas as provided in NSD are significantly easier to decode than MEG signals from THINGS-MEG.

	Low-level				High-level		
Dataset	PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	SwAV ↓
NSD (fMRI)	0.305	0.366	0.962	0.977	0.910	0.917	0.410
THINGS-MEG (per-trial average)	0.079	0.329	0.718	0.823	0.674	0.765	0.595
THINGS-MEG (per-subject average)	0.088	0.333	0.747	0.855	0.712	0.804	0.576
THINGS-MEG (no average)	0.069	0.308	0.668	0.733	0.613	0.668	0.636

Generating images from MEG. While framing decoding as a retrieval task yields promising results, it requires the true image to be in the retrieval set – a well-posed problem which presents limited use-cases in practice. To address this issue, we trained three distinct brain modules to predict the three embeddings that we use (see Section 2.5) to generate images (Fig. 5). As confirmed by the evaluation metrics of Table 1, the generated images look relatively good, with multiple generated images sharing the correct ground-truth category. However, they appear to contain limited low-level information about the true image.

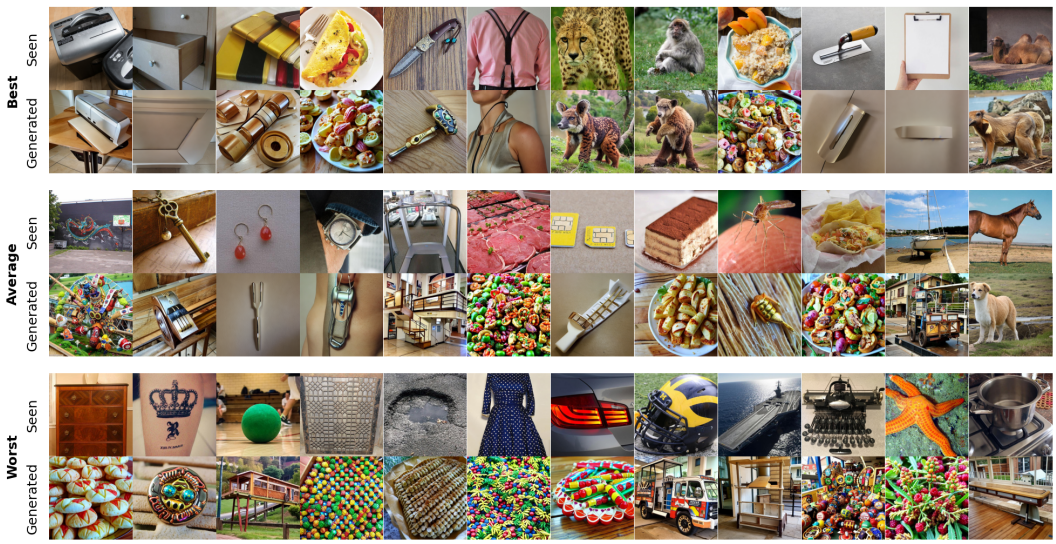

Figure 5: Examples of generated images conditioned on MEG-based latent predictions. The groups of three stacked rows represent best, average and worst generations, as evaluated by the sum of (minus) SwAV and SSIM.

The application of a very similar pipeline on an analogous fMRI dataset (Allen et al., 2022; Ozcelik & VanRullen, 2023) – using a simple Ridge regression – shows image reconstructions that share both high-level and low-level features with the true image Fig. S3). Together, these results suggest that it is not the reconstruction pipeline which fails to reconstruct low-level features, but rather the MEG signals which contain little information at that level.

4 DISCUSSION

Related work. The present study shares several elements with previous MEG and electroencephalography (EEG) studies designed not to maximize decoding performance but to understand the cascade of visual processes in the brain. In particular, previous studies have trained linear models to either (1) classify a small set of images from brain activity (Grootswagers et al., 2019; King & Wyart, 2021), (2) predict brain activity from the latent representations of the images (Cichy et al., 2017) or (3) quantify the similarity between these two modalities with representational similarity analysis (RSA) (Cichy et al., 2017; Bankson et al., 2018; Grootswagers et al., 2019; Gifford et al., 2022). While these studies also make use of image embeddings, their linear decoders are limited to classifying a small set of object classes, or to distinguishing pairs of images.

In addition, several deep neural networks have been introduced to maximize the classification of speech (Défossez et al., 2022), mental load (Jiao et al., 2018) and images (Palazzo et al., 2020; McCartney et al., 2022; Bagchi & Bathula, 2022) from EEG recordings. In particular, Palazzo et al. (2020) introduced a deep convolutional neural network to classify natural images from EEG signals. However, the experimental protocol consisted of presenting all of the images of the same class within a single continuous block, which risks allowing the decoder to rely on autocorrelated noise, rather than informative brain activity patterns (Li et al., 2020). In any case, these EEG studies focus on the categorization of a relatively small number of images classes.

In sum, there is, to our knowledge, no MEG decoding study that learns end-to-end to reliably generate an open set of images.

Impact. The present work has both fundamental and practical impacts. First, the ability to decode complex perceptual representations as a function of time promises to greatly facilitate our understanding of the processes at stake during visual processing in the brain. There is considerable work inspecting the nature and the timing of the representations built along the visual system. However, these results can be challenging to interpret, especially for high-level features. Generative decoding, on the contrary, provides concrete and, thus, interpretable predictions. Second, the most obvious use-case of brain decoding technology is to assist patients whose brain lesions challenge communication. This use-case, however, requires real-time decoding, and thus limit the use of neuroimaging modalities with low temporal resolution such as fMRI. The present effort thus paves the way to achieve this long-awaited goal.

Limitations. Our analyses highlight three main limitations to the decoding of images from MEG signals. First, the decoding of high-level semantic features prevails over low-level features: in particular, the generated images preserve semantics (*e.g.* object categories) much better than low-level features (*e.g.* contours, shading). This phenomenon is difficult to attribute to our pipeline: indeed, applying a similar procedure to 7T fMRI recordings achieves reasonably high reconstruction of low-level features (Fig. S3). Rather, this result resonates with the fact that the spatial resolution of MEG ($\approx$ cm) is much lower than 7T fMRI's ($\approx$ mm). Second, the present approach directly depends on the pretraining of several models, and only learns end-to-end to align the MEG signals to these pretrained embeddings. Our results show that this approach leads to better performance than classical computer vision features such as color histograms, fast-Fourier transforms and histogram of oriented gradients (HOG). This is consistent with a recent MEG study by Défossez et al. (2022) which showed, in the context of speech decoding, that pretrained embeddings outperformed a fully end-to-end approach. Nevertheless, it remains to be tested whether (1) fine-tuning the image and generation modules and (2) combining the different types of visual features could improve decoding performance.

Ethical implications. While the decoding of brain activity promises to help a variety of brain-lesioned patients (Metzger et al., 2023; Moses et al., 2021; Défossez et al., 2022; Liu et al., 2023; Willett et al., 2023), the rapid advances of this technology raise several ethical considerations, and most notably, the necessity to preserve mental privacy. Several empirical findings are relevant to this issue. Firstly, the decoding performance obtained with non-invasive recordings is only high for *perceptual* tasks. By contrast, decoding accuracy considerably diminishes when individuals are tasked to imagine representations (Horikawa & Kamitani, 2017; Tang et al., 2023). Second, decoding performance seems to be severely compromised when participants are engaged in disruptive tasks, such

as counting backward (Tang et al., 2023). In other words, the subjects' consent is not only a legal but also and primarily a technical requirement for brain decoding. To delve into these issues effectively, we endorse the open and peer-reviewed research standards.

Conclusion. Overall, these results provide an important step towards the decoding of the visual processes continuously unfolding in the human brain.

REFERENCES

Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022.

Subhranil Bagchi and Deepti R Bathula. EEG-ConvTransformer for single-trial EEG-based visual stimulus classification. *Pattern Recognition*, 129:108757, 2022.

Andrea Banino, Caswell Barry, Benigno Uria, Charles Blundell, Timothy Lillicrap, Piotr Mirowski, Alexander Pritzel, Martin J Chadwick, Thomas Degris, Joseph Modayil, et al. Vector-based navigation using grid-like representations in artificial agents. *Nature*, 557(7705):429–433, 2018.

B.B. Bankson, M.N. Hebart, I.I.A. Groen, and C.I. Baker. The temporal evolution of conceptual object representations revealed through models of behavior, semantics and deep neural networks. *NeuroImage*, 178:172–182, 2018. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2018.05.037. URL https://www.sciencedirect.com/science/article/pii/S1053811918304440.

G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature human behaviour*, 7(3):430–441, 2023.

Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, and Aude Oliva. Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. *NeuroImage*, 153:346–358, 2017.

Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. Decoding speech from non-invasive brain recordings. *arXiv preprint arXiv:2208.12266*, 2022.

Matteo Ferrante, Tommaso Boccato, and Nicola Toschi. Semantic brain decoding: from fMRI to conceptually similar image reconstruction of visual stimuli. *arXiv preprint arXiv:2212.06726*, 2022.

Alessandro T Gifford, Kshitij Dwivedi, Gemma Roig, and Radoslaw M Cichy. A large and rich EEG dataset for modeling human visual object recognition. *NeuroImage*, 264:119754, 2022.

Tijl Grootswagers, Amanda K Robinson, and Thomas A Carlson. The representational dynamics of visual objects in rapid serial visual processing streams. *NeuroImage*, 188:668–679, 2019.

Sébastien B Hausmann, Alessandro Marin Vargas, Alexander Mathis, and Mackenzie W Mathis. Measuring and modeling the motor system with machine learning. *Current opinion in neurobiology*, 70:11–23, 2021.

Martin N Hebart, Adam H Dickter, Alexis Kidder, Wan Y Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I Baker. THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PloS one*, 14(10):e0223792, 2019.

Martin N Hebart, Oliver Contier, Lina Teichmann, Adam H Rockter, Charles Y Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I Baker. THINGS-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *eLife*, 12:e82580, feb 2023. ISSN 2050-084X. doi: 10.7554/eLife.82580. URL https://doi.org/10.7554/eLife.82580.

Tomoyasu Horikawa and Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature communications*, 8(1):15037, 2017.

David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106, 1962.

Vinay Jayaram and Alexandre Barachant. MOABB: trustworthy algorithm benchmarking for bcis. *Journal of neural engineering*, 15(6):066011, 2018.

Zhicheng Jiao, Xinbo Gao, Ying Wang, Jie Li, and Haojun Xu. Deep convolutional neural networks for mental load classification based on EEG data. *Pattern Recognition*, 76:582–595, 2018.

Yukiyasu Kamitani and Frank Tong. Decoding the visual and subjective contents of the human brain. *Nature neuroscience*, 8(5):679–685, 2005.

Nancy Kanwisher, Josh McDermott, and Marvin M Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of neuroscience*, 17(11):4302–4311, 1997.

Jean-Rémi King and Valentin Wyart. The human brain encodes a chronicle of visual events at each instant of time through the multiplexing of traveling waves. *Journal of Neuroscience*, 41(34): 7224–7233, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Ren Li, Jared S Johansen, Hamad Ahmed, Thomas V Ilyevsky, Ronnie B Wilbur, Hari M Bharadwaj, and Jeffrey Mark Siskind. The perils and pitfalls of block design for EEG classification experiments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):316–333, 2020.

Yan Liu, Zehao Zhao, Minpeng Xu, Haiqing Yu, Yanming Zhu, Jie Zhang, Linghao Bu, Xiaoluo Zhang, Junfeng Lu, Yuanning Li, et al. Decoding and synthesizing tonal language speech from brain activity. *Science Advances*, 9(23):eadh0478, 2023.

Weijian Mai and Zhijun Zhang. Unibrain: Unify image reconstruction and captioning all in one diffusion model from human brain activity. *arXiv preprint arXiv:2308.07428*, 2023.

Ben McCartney, Barry Devereux, and Jesus Martinez-del Rincon. A zero-shot deep metric learning approach to brain–computer interfaces for image retrieval. *Knowledge-Based Systems*, 246: 108556, 2022.

Johannes Mehrer, Courtney J Spoerer, Emer C Jones, Nikolaus Kriegeskorte, and Tim C Kietzmann. An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences*, 118(8):e2011417118, 2021.

Sean L Metzger, Kaylo T Littlejohn, Alexander B Silva, David A Moses, Margaret P Seaton, Ran Wang, Maximilian E Dougherty, Jessie R Liu, Peter Wu, Michael A Berger, et al. A high-performance neuroprosthesis for speech decoding and avatar control. *Nature*, pp. 1–10, 2023.

David A Moses, Sean L Metzger, Jessie R Liu, Gopala K Anumanchipalli, Joseph G Makin, Pengfei F Sun, Josh Chartier, Maximilian E Dougherty, Patricia M Liu, Gary M Abrams, et al. Neuroprosthesis for decoding speech in a paralyzed person with anarthria. *New England Journal of Medicine*, 385(3):217–227, 2021.

Shinji Nishimoto, An T Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current biology*, 21(19):1641–1646, 2011.

John O'Keefe and Lynn Nadel. The hippocampus as a cognitive map. *Behavioral and Brain Sciences*, 2(4):487–494, 1979.

Furkan Ozcelik and Rufin VanRullen. Brain-diffuser: Natural scene reconstruction from fMRI signals using generative latent diffusion. *arXiv preprint arXiv:2303.05334*, 2023.

Simone Palazzo, Concetto Spampinato, Isaak Kavasidis, Daniela Giordano, Joseph Schmidt, and Mubarak Shah. Decoding brain representations by multimodal learning of neural activity and visual features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3833–3849, 2020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H Falk, and Jocelyn Faubert. Deep learning-based electroencephalography analysis: a systematic review. *Journal of neural engineering*, 16(5):051001, 2019.

Martin Schrimpf, Idan Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua Tenenbaum, and Evelina Fedorenko. Artificial neural networks accurately predict language processing in the brain. *BioRxiv*, pp. 2020–06, 2020.

Paul S Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Ethan Cohen, Aidan J Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, et al. Reconstructing the mind's eye: fMRI-to-image with contrastive learning and diffusion priors. *arXiv preprint arXiv:2305.18274*, 2023.

Katja Seeliger, Umut Güçlü, Luca Ambrogioni, Yagmur Güçlütürk, and Marcel AJ van Gerven. Generative adversarial networks for reconstructing natural images from brain activity. *NeuroImage*, 181:775–785, 2018.

Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. *bioRxiv*, 2023. doi: 10.1101/2022.11.18.517004. URL `https://www.biorxiv.org/content/early/2023/03/11/2022.11.18.517004`.

Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, pp. 1–9, 2023.

Armin Thomas, Christopher Ré, and Russell Poldrack. Self-supervised learning of brain dynamics from broad neuroimaging data. *Advances in Neural Information Processing Systems*, 35:21255–21269, 2022.

Stefan Van der Walt, Johannes L Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. scikit-image: image processing in python. *PeerJ*, 2:e453, 2014.

Rufin VanRullen and Leila Reddy. Reconstructing faces from fMRI patterns using deep generative neural networks. *Communications biology*, 2(1):193, 2019.

Francis R Willett, Erin M Kunz, Chaofei Fan, Donald T Avansino, Guy H Wilson, Eun Young Choi, Foram Kamdar, Matthew F Glasser, Leigh R Hochberg, Shaul Druckmann, et al. A high-performance speech neuroprosthesis. *Nature*, pp. 1–6, 2023.

Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.

Bohan Zeng, Shanglin Li, Xuhui Liu, Sicheng Gao, Xiaolong Jiang, Xu Tang, Yao Hu, Jianzhuang Liu, and Baochang Zhang. Controllable mind visual diffusion model. *arXiv preprint arXiv:2305.10135*, 2023.

A APPENDIX

A.1 LINEAR RIDGE REGRESSION SCORES ON PRETRAINED IMAGE REPRESENTATIONS

We provide a (5-fold cross-validated) Ridge regression baseline (Table S1) for comparison with our brain module results of Section 3, showing considerable improvements for the latter.

Table S1: Image retrieval performance of a linear Ridge regression baseline on pretrained image representations

		Top-5 acc (%) ↑		**Median relative rank ↓**	
Latent kind	Latent name	Small set	Large set	Small set	Large set
Text/Image alignment	CLIP-Vision (CLS)	10.5	0.50	0.23	0.34
	CLIP-Text (mean)	6.0	0.25	0.42	0.43
	CLIP-Vision (mean)	5.5	0.46	0.32	0.37
Feature engineering	Color histogram	7.0	0.33	0.31	0.40
	Local binary patterns (LBP)	3.5	0.37	0.34	0.44
	FFT 2D (as real)	4.5	0.46	0.40	0.45
	HOG	3.0	0.42	0.45	0.46
	FFT 2D (log-PSD and angle)	2.0	0.37	0.47	0.46
Variational autoencoder	AutoKL	7.5	0.54	0.24	0.38
	VDVAE	8.0	0.50	0.33	0.43
Self-supervised learning	DINOv2 (CLS)	7.5	0.46	0.25	0.35
Supervised	VGG-19	**12.5**	**1.04**	**0.18**	**0.33**
	ResNet-101	4.0	0.37	0.36	0.42
	DenseNet-201	5.0	0.29	0.39	0.45
	Wide ResNet-101-2	3.5	0.42	0.40	0.46
	MobileNet v3	3.5	0.42	0.40	0.42

A.2 HYPERPARAMETER SEARCH

We run a hyperparameter search to find an appropriate configuration (MEG preprocessing, optimizer, brain module architecture and loss definition) for the MEG-to-image retrieval task ($\lambda = 0$). We randomly split the 79,392 (MEG, image) pairs of the adapted training set (Section 2.8) into 60%-20%-20% train, valid and test splits such that all presentations of a given image are contained in the same split. We use the validation split to perform early stopping and the test split to evaluate the performance of a configuration.

For the purpose of this search we pick CLIP-Vision (CLS) latent as a representative latent, since it achieved good retrieval performance in preliminary experiments. We run the search six times using two different random seed initializations for the brain module and three different random train/valid/test splits. Fig. S1 summarizes the results of this hyperparameter search.

Based on this search, we use the following configuration: MEG window (t_{min}, t_{max}) of $[-0.5, 1.0]$ s, learning rate of 3×10^{-4}, batch size of 128, brain module with two convolutional blocks and both the spatial attention and subject layers of Défossez et al. (2022), affine projection temporal aggregation layer with a single block in the CLIP projection head, and full CLIP loss (including learned temperature parameter, normalization along both axes and symmetric terms). The final architecture configuration is presented in Table S2.

A.3 FULL-WINDOW MEG-BASED IMAGE RETRIEVALS

Fig. S2 shows examples of retrieved images based on the best performing latents identified in Section 3.

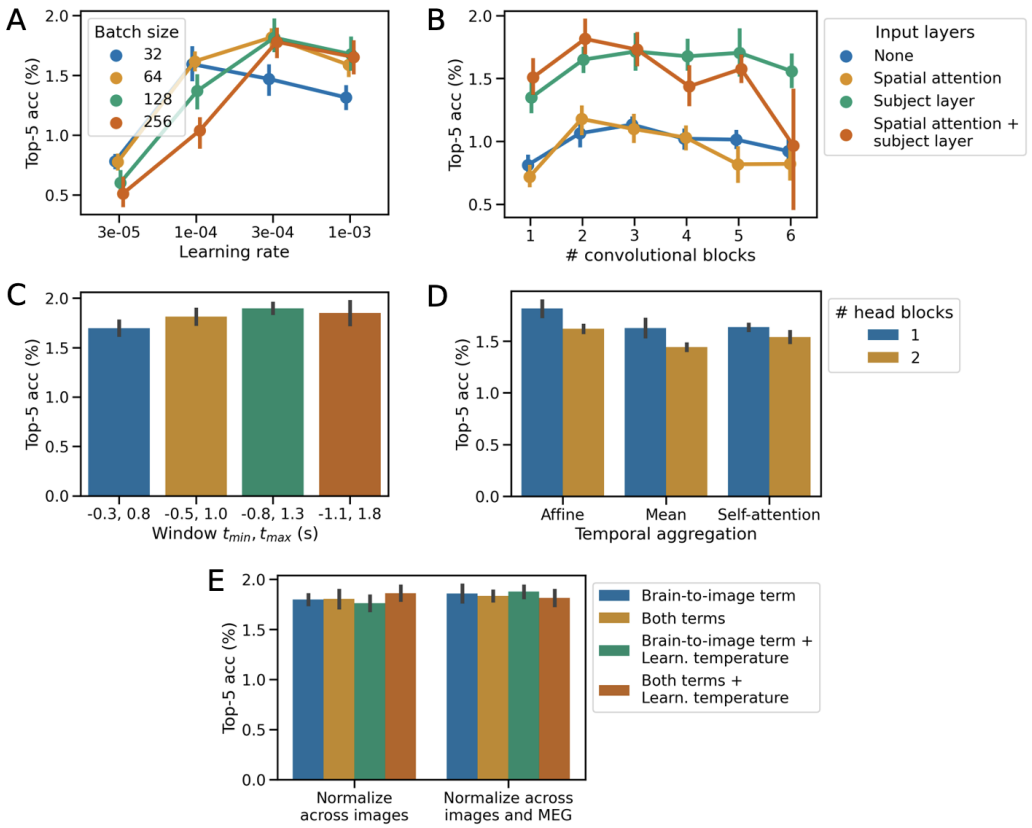

Figure S1: Hyperparameter search results for the MEG-to-image retrieval task, presenting the impact of (**A**) optimizer learning rate and batch size, (**B**) number of convolutional blocks and use of spatial attention and/or subject-specific layers in the brain module, (**C**) MEG window parameters, (**D**) type of temporal aggregation layer and number of blocks in the CLIP projection head of the brain module, and (**E**) CLIP loss configuration (normalization axes, use of learned temperature parameter and use of symmetric terms). Chance-level performance top-5 accuracy is 0.05%.

Table S2: Brain module configuration adapted from Défossez et al. (2022) for use with a target latent of size 768 (*e.g.* CLIP-Vision (CLS), see Section 2.4) in retrieval settings.

Layer	**Input shape**	**Output shape**	**# parameters**
Spatial attention block	(272, 181)	(270, 181)	552,960
Linear projection	(270, 181)	(270, 181)	73,170
Subject-specific linear layer	(270, 181)	(270, 181)	291,600
Residual dilated conv block 1	(270, 181)	(320, 181)	1,183,360
Residual dilated conv block 2	(320, 181)	(320, 181)	1,231,360
Linear projection	(320, 181)	(2048, 181)	1,518,208
Temporal aggregation	(2048, 181)	(2048, 1)	182
MLP projector	(2048, 1)	(768, 1)	1,573,632
Total			6,424,472

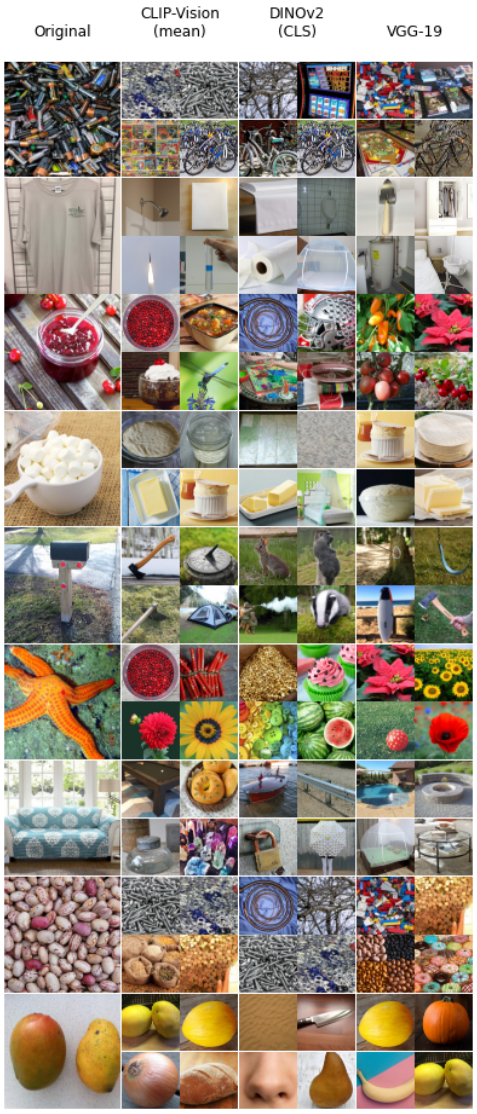

Figure S2: Representative examples of retrievals (top-4) using models trained on full windows (from -0.5 s to 1 s after image onset). Retrieval set: $N = 6,059$ images from $1,196$ categories.

A.4 IMAGE EMBEDDINGS

We evaluate the performance of linear baselines and of a deep convolutional neural network on the MEG-to-image retrieval task using a set of classic visual embeddings. We grouped these embeddings by their corresponding paradigm:

Supervised learning. DenseNet-121,DenseNet-169,DenseNet-201, MobileNet v2, MobileNet v3, ResNet-101, ResNet-18, ResNet-50, ResNext101-32-8d, ResNext50-32-4d, VGG-16,VGG-19, Wide ResNet-101-2, Wide ResNet-50-2.

Text/Image alignment. CLIP-Vision, CLIP-Text, and their CLS and MEAN pooling.

Self-supervised learning. DINOv1, DINOv2 and their CLS and MEAN pooling.

Figure S3: Examples of generated images conditioned on fMRI-based latent predictions. The groups of three stacked rows represent best, average and worst retrievals, as evaluated by the sum of (minus) SwAV and SSIM.

Variational autoencoders. The activations of the 31 first layers of the very deep variational-autoencoder (VDVAE), and the Kullback-Leibler variational-autoencoder (AutoKL) used in the generative module (Section 2.5).

Engineered features. The color histogram of the seen image (8 bins per channels); the local binary patterns (LBP) using the implementation in OpenCV 2 (Bradski, 2000) with 'uniform' method, $P = 8$ and $R = 1$; the Histogram of Oriented Gradients (HOG) using the implementation of sk-image (Van der Walt et al., 2014) with 8 orientations, 8 pixels-per-cell and 2 cells-per-block.

A.5 7T FMRI DATASET

The Natural Scenes Dataset (NSD) (Allen et al., 2022) contains fMRI data from 8 participants viewing a total of 73,000 RGB images. It has been successfully used for reconstructing seen images from fMRI in several studies (Takagi & Nishimoto, 2023; Ozcelik & VanRullen, 2023; Scotti et al., 2023). In particular, these studies use a highly preprocessed, compact version of fMRI data ("betas") obtained through generalized linear models fitted across multiple repetitions of the same image.

Each participant saw a total of 10,000 unique images (repeated 3 times each) across 37 sessions. Each session consisted in 12 runs of 5 minutes each, where each image was seen during 3 s, with a 1-s blank interval between two successive image presentations. Among the 8 participants, only 4 (namely 1, 2, 5 and 7) completed all sessions.

To compute the three latents used to reconstruct the seen images from fMRI data (as described in Section 2.5) we follow Ozcelik & VanRullen (2023) and train and evaluate three distinct Ridge regression models using the exact same split. That is, for each of the four remaining participants, the 9,000 uniquely-seen-per-participant images (and their three repetitions) are used for training, and a common set of 1000 images seen by all participant is kept for evaluation (also with their three repetitions). We report reconstructions and metrics for participant 1.

The α coefficient for the $L2$-regularization of the regressions are cross-validated with a 5-fold scheme on the training set of each subject. We follow the same standardization scheme for inputs and predictions as in (Ozcelik & VanRullen, 2023).

Fig. S3 presents generated images obtained using the NSD dataset (Allen et al., 2022).