LATENT PREDICTION YIELDS CROSS-SUBJECT ALIGNABLE REPRESENTATIONS IN SENSORY-EVOKED FMRI

Anonymous authors
Paper under double-blind review

ABSTRACT

Decoding sensory-evoked functional magnetic resonance imaging (fMRI) data is a central analytical tool in cognitive neuroscience. Here we ask whether self-supervised learning (SSL), which has been scarcely explored in this domain, can yield representations that support strong downstream performance. To do so, we adapt a joint embedding predictive architecture (JEPA) to train participant-specific encoders on brain data alone. The encoders never see the stimulus images or any representation derived from them. Applied to the Natural Scenes Dataset (NSD), we evaluate our learned brain embeddings on the downstream tasks of brain captioning and image reconstruction, as well as cross-participant generalization. Our self-supervised approach achieves competitive performance with parameter-rich supervised pipelines that are trained end-to-end with access to image and language data. We further find that encoders trained independently for different participants can be aligned fully unsupervised, using neither images nor stimuli shared between them. After alignment, a neural decoding head trained on one participant can be applied to another without fine-tuning and retains most of its within-subject performance on image reconstruction and brain captioning. By removing the need for images and shared stimuli, our approach is a step towards pooling fMRI data across participants and, eventually, across datasets.

1 INTRODUCTION

The decoding of information from brain activity is a central analytical tool in cognitive neuroscience. Classical applications ask whether experimental conditions, visual categories, or other stimulus properties can be read out from measured brain responses alone. More recently, advances in generative modeling and the availability of large-scale neural datasets, such as the Natural Scenes Dataset (NSD) (Allen et al., 2022), have substantially expanded the scope of which information can be derived from brain data. Capabilities of brain decoding now include largely veridical image reconstruction from functional MRI (fMRI) responses, obtained while participants observe visual stimuli (Takagi & Nishimoto, 2023; Ozcelik & van Rullen, 2023; Scotti et al., 2024), up to approaches that map brain activity into language representations, enabling versatile brain captioning and question-answering interfaces beyond reconstruction itself (Mai & Zhang, 2023; Bosch et al., 2025).

These approaches, however, place substantial demands on the available data. Modern decoding pipelines are typically trained separately for each participant, and many additionally rely on paired supervision between brain responses and stimulus-derived representations, such as image embeddings or language annotations. Both requirements are restrictive. Even unusually large fMRI datasets contain orders of magnitude fewer training examples per participant than are commonly used to train modern machine-learning models, simply because acquiring fMRI data is time-consuming and expensive. Pooling data across participants could substantially increase the amount of brain data available for representation learning, but this is difficult because brains differ in functional organization and in the precise spatial pattern through which information is represented. Consequently, a decoder trained on one participant does not straightforwardly generalize to another. Existing approaches to functional alignment can reduce this problem, but typically depend on paired measurements, i.e., responses from multiple participants to the same stimuli.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

An appealing alternative would therefore be to learn brain representations that address both limitations at once: such a representation should make effective use of the brain data itself, without requiring large-scale external supervision during representation learning, while at the same time enabling versatile downstream applications and exposing a structure that can be aligned across participants despite their anatomical and functional differences. Self-supervised learning (SSL) provides a natural framework for this objective. Rather than learning brain representations by predicting an image embedding or a language annotation, we learn them from the structure that the measured brain responses share across repetitions of the same stimulus. Because the objective is self-supervised, the resulting representation is not tied to the feature space of a particular external image or language model, and can serve multiple readouts. In addition, the objective does not depend on the stimulus modality: it applies to any experiment that repeats its stimuli.

Here, we explore this possibility for stimulus-evoked fMRI using a joint-embedding predictive architecture (JEPA). SSL has previously been applied to fMRI, particularly in the resting-state setting for downstream phenotype or clinical prediction (Ortega Caro et al., 2023; Dong et al., 2024). We instead ask whether it can recover representations of stimulus-evoked activity that retain the fine-grained information required for generative decoding. We pre-train one encoder per participant on NSD, asking it to predict the latent representation of masked cortical parcels from a different repetition of the same stimulus. Pairing repetitions this way leaves stimulus-locked structure as the only predictable signal and does not reward the model for measurement-specific noise; the model also never sees the stimulus, an image embedding, or a language annotation during pre-training.

We find that the resulting brain embeddings are highly expressive. After self-supervised pre-training, we freeze the encoders and train downstream heads for image reconstruction and brain captioning using MindEye2 (Scotti et al., 2024) and CorText (Bosch et al., 2025), respectively. Despite the absence of image or language supervision during encoder training, these brain embeddings support downstream performance competitive with the corresponding pipelines trained end-to-end with stimulus supervision. Thus, much of the information required for mapping brain activity into modern image and language models can be captured through latent prediction directly within brain data, reducing the amount of task-specific parameter learning required at the decoding stage.

Self-supervised training also yields representations that can be aligned across participants without externally defined anchors, such as shared stimuli or stimulus content. This has recently been shown for compact, largely linear embeddings of NSD participants (Marcos-Manchón et al., 2026). What remains open is whether alignability survives in representations rich enough to drive generative decoders and whether a decoder trained on one participant works on the aligned representations of another. Applying that alignment procedure, extended to parcel sequences, to our per-participant encoders, we find the full parcel embeddings still related by orthogonal maps recoverable without paired data, and a decoding head trained on one participant applies unchanged to another's aligned representations, retaining most of its within-participant quality. Self-supervised pre-training therefore provides both an expressive within-participant representation and a basis for cross-participant transfer.

Our contributions are:

- We introduce self-supervised pre-training of stimulus-evoked fMRI encoders using a brain-only JEPA objective. The objective exploits repeated responses to the same stimulus, but requires no access to the stimulus itself or to externally pre-trained image or language representations.
- We show that the resulting frozen brain embeddings simultaneously support both image reconstruction and brain captioning at performance levels competitive with corresponding end-to-end supervised pipelines, despite requiring substantially less task-specific parameter learning.
- We show that independently trained, participant-specific brain embeddings can be aligned fully unsupervised, without images or stimuli shared across participants, enabling cross-participant decoding without retraining the downstream decoder.

Together, these results suggest that latent prediction can provide a common representational basis for both expressive brain decoding and cross-participant generalization. By decoupling representation learning from stimulus-derived supervision and cross-participant alignment from shared stimuli,

this approach opens a path towards combining brain data across participants and, ultimately, across datasets that were not originally acquired with joint decoding in mind.

2 Related Work

Self-Supervised Representation Learning. SSL is a family of training objectives that require only input data and no target labels. It is commonly used to pre-train representations that can be deployed across a range of downstream tasks. Prominent SSL objectives include reconstruction or masked prediction in the input space (Vincent et al., 2008; Pathak et al., 2016; Devlin et al., 2019; He et al., 2022; Xie et al., 2022), and prediction or matching in latent representation spaces (Chen et al., 2020; Grill et al., 2020; Caron et al., 2021; Bardes et al., 2022). Joint embedding predictive architectures (JEPAs) (LeCun, 2022; Assran et al., 2023; Bardes et al., 2024; Assran et al., 2025) are a more recent class of latent-prediction models: rather than reconstructing raw observations, they predict latent representations of masked regions from a visible context. Because unpredictable input-level details need not be represented faithfully (Littwin et al., 2024), this objective may be particularly suitable for fMRI, where scanner, physiological, and other non-stimulus-related variation is common.

SSL methods have been widely applied in neuroimaging (Kim et al., 2023; Ortega Caro et al., 2023; Yang et al., 2024; Wang et al., 2025; 2026a;b), though many operate on resting-state data and are evaluated on phenotype and clinical readouts. In that setting, Dong et al. (2024) report that a JEPA objective outperforms reconstruction- and contrastive-based SSL on demographic and disease prediction, motivating our use of this objective family for sensory-evoked fMRI. Closest to our setting, Marcos-Manchón et al. (2026) construct subject-specific representations on NSD (Allen et al., 2022) using a largely linear pipeline (PCA and multiset CCA), refine them with a contrastive objective, and evaluate them through cross-trial retrieval. We differ by learning these representations end-to-end and evaluating them on generative readouts rather than retrieval alone.

Decoding from fMRI: vision and language. Advances in generative architectures for vision and language have enabled increasingly detailed decoding of sensory-evoked fMRI. One line of work maps brain signals into the latent space of generative image models to reconstruct the viewed stimulus (Takagi & Nishimoto, 2023; Ozcelik & van Rullen, 2023; Chen et al., 2023b; Scotti et al., 2023); a parallel line maps them into language models to generate textual descriptions instead (Ferrante et al., 2023; Mai & Zhang, 2023; Chen et al., 2023a; Huang, 2024; Qiu et al., 2025). We adopt MindEye2 (Scotti et al., 2024) for image reconstruction and CorText (Bosch et al., 2025) for captioning. Both methods train the brain-side encoder against stimulus-derived targets, such as frozen image embeddings or caption tokens. We instead use these pipelines as downstream readouts — trained against the same targets — of a frozen brain encoder whose pre-training objective has no access to stimulus images or stimulus-derived representations.

Cross-Subject Functional Alignment. Brains differ across individuals in both anatomy and functional organization, so the same measurement location does not correspond across people. Because establishing that correspondence is difficult, decoding models are typically built separately for each individual. Hyperalignment addressed this by learning transformations from responses to shared stimuli (Haxby et al., 2011), an approach extended by the shared response model (Chen et al., 2015) and optimal-transport formulations (Bazeille et al., 2019) and inherited by several deep learning-based methods (Chen et al., 2016; Huang et al., 2021). More recent methods remove the requirement for shared stimuli, but instead anchor participants to representations of pre-trained image encoders (Wasserman et al., 2026; Papathanasiou et al., 2026). Our setting removes both requirements: neither pre-training nor alignment uses shared stimuli or representations derived from the stimulus images.

Outside neuroimaging, it has already been shown that different embedding spaces can be brought into correspondence without paired examples by matching their internal similarity structure (Conneau et al., 2018; Alvarez-Melis & Jaakkola, 2018; Jha et al., 2025). Marcos-Manchón et al. (2026) recently adopted this framework for sensory-evoked fMRI data, aligning per-participant embeddings without paired stimuli across participants or access to stimulus images. We use their alignment procedure to test whether independently and end-to-end learned neural representations develop corresponding geometries.

3 METHOD

Our method has three components: we pre-train one JEPA encoder per participant on parcellated sensory-evoked fMRI responses (Section 3.2), evaluate the resulting brain embeddings with generative decoding heads trained on top of the frozen encoder (Section 3.3), and align the independently trained encoders across participants without shared stimuli (Section 3.4). Appendix A lists the parameters for all three. We first describe the data and the parcel representation that all three build on.

3.1 DATASET

We use NSD (Allen et al., 2022) for both pre-training and downstream evaluation. NSD is a sensory-evoked 7T fMRI dataset in which eight participants viewed up to 10,000 images from MS COCO (Lin et al., 2014) over 40 scanning sessions, where each image was presented up to three times in total. We restrict our experiments to participants 1, 2, 5, and 7, the four who completed all scanning sessions. For each participant, the training set comprises the 9,000 unique stimuli seen only by that participant, while the 1,000 shared stimuli, which all participants viewed, are held out for evaluation. We use the fsaverage-surface betas provided with NSD and z-score each session independently within each participant, with train and test sets standardized separately. We then parcellate the surface data with the 500-parcel 7-network Schaefer atlas (Schaefer et al., 2018), yielding 500 parcels per trial.

We explore restricting the readout to a subset of parcels containing visually responsive regions. We define this subset anatomically and independently of any functional response. Glasser et al. (2016) group their 180 cortical areas into 22 coarser cortical divisions; we assign each Schaefer parcel to the division containing the largest share of its vertices on the shared fsaverage surface, and retain the parcel tokens assigned to the five visual divisions — Primary Visual, Early Visual, Dorsal Stream Visual, Ventral Stream Visual, and MT+ Complex. This yields 66 parcels spanning 41,248 of the 299,761 labeled vertices (13.8%), 35 in the left and 31 in the right hemisphere. Because the two atlases' boundaries do not coincide, individual Schaefer parcels may span multiple Glasser divisions; nonetheless, 44 of the 66 selected parcels lie entirely within visual divisions, and, on average, 93.1% of a selected parcel's vertices do. Since the criterion uses only anatomical atlas overlap, the subset is identical across participants and involves no selection on the neural data.

3.2 SELF-SUPERVISED BRAIN ENCODER

The functional parcellation described above maps naturally onto the JEPA setup, which learns by predicting the representations of masked tokens — typically image or video patches, and here the parcels of a single trial's brain response. Figure 1 gives an overview of the training process.

Let $x \in \mathbb{R}^V$ denote the response of a single trial, where V is the number of fsaverage vertices covered by the atlas. The parcellation partitions these vertices into $P = 500$ disjoint parcels, so that $x = (x_1, \ldots, x_P)$ with $x_p \in \mathbb{R}^{V_p}$ and $\sum_{p=1}^{P} V_p = V$. Within each parcel, vertices are weighted by their reliability — how consistently a vertex responds across repeated presentations of the same stimulus — so that noise-dominated vertices contribute less: $\tilde{x}_p = w_p \odot x_p$, with $w_p \in \mathbb{R}^{V_p}$ computed following Marcos-Manchón et al. (2026) and normalized to sum to one.

The context encoder f_θ begins with a parcel embedding layer. Because the number of vertices V_p differs across parcels, it comprises a two-layer MLP $e_p : \mathbb{R}^{V_p} \to \mathbb{R}^D$ per parcel. Together with learnable positional embeddings $E \in \mathbb{R}^{P \times D}$, it maps a brain response of V vertices to a sequence of P tokens of width D. We use learnable rather than structured positional embeddings, as we do not want to impose a neighborhood structure on the parcel sequence. A stack of transformer blocks then operates on this sequence. The predictor g_ϕ is a transformer at width D_p and predicts masked tokens using learnable mask tokens. Following the standard JEPA procedure, we initialize a target encoder $f_{\bar{\theta}}$ from the context encoder and update it as an exponential moving average of its weights.

Unlike image patches, contiguous blocks of the parcel sequence do not correspond to anatomically neighboring cortical regions, so we sample random token sets rather than the contiguous blocks commonly used in JEPA pre-training. Masking randomly samples one context set $C \subset \{1, \ldots, P\}$

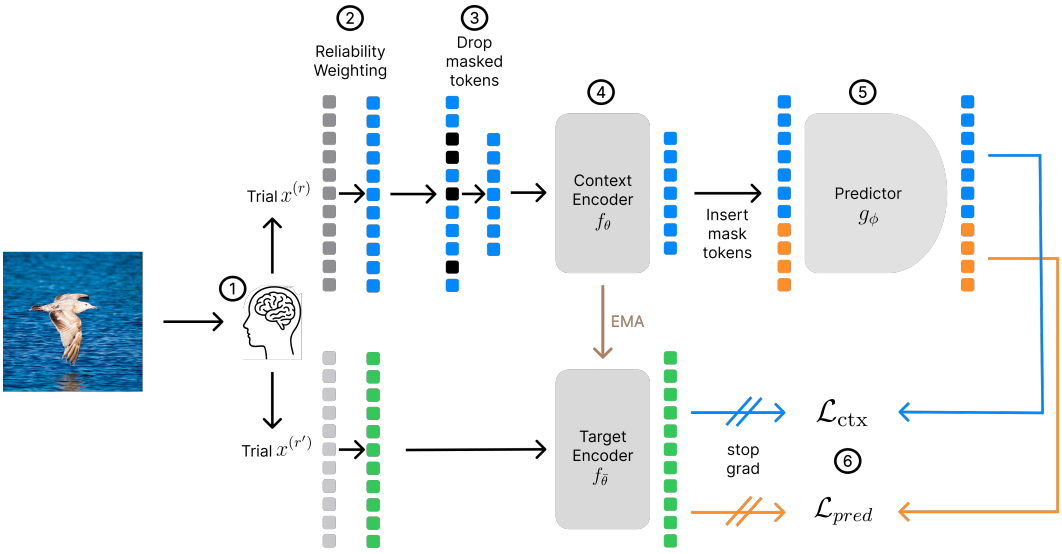

Figure 1: Overview of our self-supervised brain encoder training. **(1)** A subject's brain response to a stimulus image is recorded across repeated presentations; we sample two of the three repetitions, r and r', one for the context encoder path and one for the target encoder path. After Schaefer parcellation, each trial's brain response is represented as a sequence of parcels. **(2)** Reliability weights are applied to the raw vertices within each parcel. **(3)** Mask positions are sampled, and the corresponding tokens are dropped from the context encoder input sequence. **(4)** The context encoder f_θ produces context representations using per-parcel MLPs, followed by transformer layers. The target encoder $f_{\bar{\theta}}$, updated as an exponential moving average of the context encoder, encodes the unmasked complementary trial; no gradient flows through this path. **(5)** Using learnable mask tokens, the predictor g_ϕ predicts the missing positions from the context representations. **(6)** Predictor outputs at masked positions, as well as at context positions, are supervised against the corresponding target representations under an L_2 loss.

and M target sets $T_1, \ldots, T_M \subset \{1, \ldots, P\}$. We sample context and target sets independently, and remove tokens appearing in any target set from the context, such that $C \cap T_m = \emptyset$ for all m.

To encourage the learned brain embeddings to discard stimulus-independent information, the context and target encoders receive *different repetitions of the same stimulus*. Two repetitions share the stimulus but differ in scanner noise, physiological state, and session effects. Since these components are independent across presentations, they carry no information about the target representation, and the objective therefore gives the encoder no incentive to represent them; only stimulus-related information is predictable from one repetition to the other. Let $x^{(r)}$ and $x^{(r')}$ be two repetitions of one image, sampled without replacement from the three available and resampled each epoch. Given the context representations of $\tilde{x}^{(r)}$, the predictor produces the sequence

$$\hat{z} = g_\phi\big(f_\theta(\tilde{x}^{(r)}, C),\, T_m\big) = \big[\hat{z}_C \mid \hat{z}_{T_m}\big]$$

while the target encoder produces

$$z = f_{\bar{\theta}}\big(\tilde{x}^{(r')}\big)$$

for the entire unmasked trial $\tilde{x}^{(r')}$. Following V-JEPA 2.1 (Mur-Labadia et al., 2026), we add a loss term over context predictions, making our per-sample loss

$$\mathcal{L} = \mathcal{L}_{\text{ctx}} + \mathcal{L}_{\text{pred}} = \frac{1}{|C|}\big\|\hat{z}_C - z_C\big\|_2^2 + \frac{1}{M}\sum_{m=1}^{M}\frac{1}{|T_m|}\big\|\hat{z}_{T_m} - z_{T_m}\big\|_2^2.$$

3.3 Generative Neural Decoding

To evaluate the quality of our pre-trained encoder, we train generative decoding heads on top of the resulting brain embeddings. Throughout, we keep the encoder frozen: its weights are never updated by any task-specific learning signal. All readouts use the target encoder from the pre-training stage.

Image Reconstruction. The first task is image reconstruction, where the goal is to generate the image a participant viewed while their brain response was recorded. We adopt the MindEye2 (Scotti et al., 2024) pipeline, which uses pre-trained CLIP and VAE image representations as supervision targets for its decoder. Its head consumes a 256×1664 token sequence produced by the MindEye2 backbone. To bridge into this space from our brain embeddings of shape $P \times D$, we apply an attentive pooling layer with 256 queries followed by a linear map to 1664 dimensions. The pipeline trains a contrastive projection alongside the reconstruction objective, from which we additionally report retrieval: *image* retrieval ranks the correct image among a set of candidates given a brain response, whereas *brain* retrieval ranks the correct brain response given an image. In contrast to MindEye2, we do not pre-train the head on other subjects. The inputs also differ: MindEye2 operates on the participant-specific `nsdgeneral` region, defined by the NSD authors as the voxels most responsive to visual stimuli, whereas our encoders read the Schaefer parcellation.

Brain Captioning. The second task is decoding textual descriptions of the stimulus images from the brain embeddings. We use the CorText (Bosch et al., 2025) framework, which integrates brain data into a pretrained LLM, enabling open-ended, natural-language decoding akin to visual question-answering setups. Most LLM-based methods pair a brain encoder with a vision-language model and train the encoder against the representations of that model's vision encoder, so that image representations act as supervision targets. CorText instead trains its brain encoder end-to-end from brain responses alone, under a cross-entropy loss over target answer tokens sourced from VQA datasets defined over COCO images (see Bosch et al. (2025) for more information). Our pre-trained encoders are therefore a drop-in replacement for its brain encoder, with no image supervision entering the pipeline at any point. We project our brain embeddings into the representation space of the pre-trained LLM (Llama 3.1 8B-Instruct) with a projection layer consisting of a two-layer MLP, following common practice in vision-language models (Liu et al., 2024). Training proceeds in two stages: first, only the projection layer is trained; second, the LLM is additionally tuned with low-rank adaptation (Hu et al., 2022). The pre-trained brain encoder remains frozen in both stages.

3.4 Cross-Subject Alignment

To align our independently pre-trained brain encoders across participants, we employ the alignment scheme of Marcos-Manchón et al. (2026). The method matches the internal similarity structure of two embedding spaces without paired samples: it summarizes each space with k-means centroids, matches centroids across spaces by solving a quadratic assignment problem on their similarity structure, and uses the resulting correspondence to fit an orthogonal map via Procrustes transformation. This initial map is then refined by iterative closest point, which alternates between re-estimating pseudo-pairs under the current map and refitting it. All steps use only the representations of each participant's brain responses from the training set; no stimulus correspondence or label information enters at any point.

While the pipeline of Marcos-Manchón et al. (2026) produces a single 128-d embedding per participant, our encoders produce a sequence of P tokens of width D, with $PD \gg 128$. Learning a single map $R \in \mathbb{R}^{PD \times PD}$ is infeasible, so we instead fit one orthogonal map per token, $R^{(t)} \in \mathbb{R}^{D \times D}$. We whiten the representations with each participant's own training statistics — a single ZCA transform shared across tokens rather than a per-token one — which improves alignment quality (see Table 4). Writing W_{A} and μ_{A} for participant A's whitening matrix and mean, a token t of A's brain embedding is mapped into B's frame as

$$z^{(t)}_{\mathrm{A}\to\mathrm{B}} = W_{\mathrm{B}}^{-1}\Big(W_{\mathrm{A}}\big(z^{(t)}_{\mathrm{A}} - \mu_{\mathrm{A}}\big)R^{(t)}_{\mathrm{A}\to\mathrm{B}}\Big) + \mu_{\mathrm{B}}. \tag{1}$$

Both W and μ are estimated from each participant's own training representations, so the alignment remains free of cross-participant supervision.

We evaluate alignment quality using cross-participant retrieval: given one participant's representation of a stimulus, the task is to retrieve another's representation of the same stimulus. We further evaluate cross-participant decoding: A's representations are mapped into B's frame, where B's decoding head reads them without retraining.

Table 1: Caption generation from brain activity on the NSD test set, averaged across subjects 1, 2, 5, and 7. Higher is better for all metrics; the best result in each column and group (within-subject vs. cross-subject) is shown in bold. All variants of our method use frozen pre-trained brain encoders within the CorText pipeline. *Visual readout* encodes all 500 parcels but exposes only visual cortex tokens to the decoding head, while *Full sequence* uses all 500 encoded parcels. For cross-subject decoding, a head trained on one subject is applied to the aligned representations of another, averaged over all 12 ordered subject pairs. Control conditions remove stimulus correspondence, brain input, or cross-subject alignment, respectively. † CorText is trained fully supervised.

Method	BLEU				METEOR	ROUGE-L	CIDEr	CLIP-S	RefCLIP-S
	1	2	3	4					
CorText†	**0.628**	**0.440**	0.300	0.204	**0.202**	0.462	**0.679**	**0.649**	**0.716**
Ours: within-subject									
Full sequence	0.625	0.438	0.303	**0.212**	0.201	**0.464**	0.673	0.635	0.705
Visual readout	0.623	0.439	**0.304**	**0.212**	0.200	0.463	**0.674**	0.633	0.703
Control, within-subject									
Trial permutation	0.391	0.176	0.078	0.039	0.090	0.296	0.061	0.394	0.474
Zeroed betas	0.190	0.078	0.028	0.010	0.094	0.193	0.001	0.364	0.428
Ours, cross-subject									
Full sequence	0.500	0.317	0.197	0.127	0.159	0.396	0.394	0.568	0.639
Visual readout	**0.549**	**0.359**	**0.233**	**0.155**	**0.168**	**0.416**	**0.453**	**0.579**	**0.651**
Control, cross-subject									
Visual-readout w/o alignment	0.038	0.006	0.000	0.000	0.017	0.046	0.000	0.503	0.569

Table 2: Image reconstruction from brain activity on the NSD test set, scores averaged across subjects 1, 2, 5, and 7. Arrows indicate the direction of improvement; the best result in each column and group (within-subject vs. cross-subject) is shown in bold. All variants of our method use frozen pre-trained brain encoders within the MindEye2 pipeline. *Visual readout*, *Full sequence*, the cross-subject protocol, and the control conditions are defined as in Table 1. Retrieval uses sets of 300 candidates sampled from the full test set, with scores averaged over 30 independent draws. Brain retrieval is undefined for the zeroed-betas control, where every trial yields the same embedding, and is therefore omitted. † MindEye2 is trained fully supervised and uses all other subjects to pre-train the decoder.

Method	Low-Level				High-Level				Retrieval	
	PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	Eff↓	SwAV↓	Image↑	Brain↑
MindEye2†	**0.278**	0.328	95.2	**99.0**	**96.4**	**94.5**	**0.622**	**0.343**	**98.8**	**98.3**
Ours, within-subject										
Full sequence	0.262	**0.354**	94.9	98.4	94.5	92.9	0.647	0.363	97.2	97.3
Visual readout	0.260	0.353	**95.3**	98.5	94.7	93.0	0.642	0.363	97.2	97.5
Control, within-subject										
Trial permutation	0.017	0.288	49.6	50.5	50.3	49.8	0.972	0.641	0.2	0.2
Zeroed betas	0.007	0.350	49.7	50.1	49.8	49.9	0.990	0.709	0.3	-
Ours, cross-subject										
Full sequence	0.188	0.346	**88.3**	**94.2**	**87.3**	**85.2**	**0.767**	**0.438**	**69.1**	**75.5**
Visual readout	**0.194**	**0.361**	88.0	93.5	86.2	84.0	0.777	0.447	62.4	71.0
Control, cross-subject										
Visual-readout w/o alignment	0.011	0.244	51.7	53.5	51.6	52.1	0.977	0.634	0.6	0.5

4 RESULTS

We report within-subject brain captioning and image reconstruction results in Tables 1 and 2, respectively (see Appendix B for per-subject results and Appendix C for example reconstructions). Two controls remove the information decoding depends on: a *trial permutation* control that breaks the correspondence between brain responses and stimuli, and a *zeroed betas* control that replaces the input with zeros. Our frozen pre-trained encoders outperform both controls by a wide margin on

Table 3: Pre-training ablation, measured by within-subject cross-trial retrieval. One repetition of a stimulus is ranked by cosine similarity against the embeddings of another repetition over the full 1000-stimulus test set, with no trained decoder; chance is mean rank 500.5 and Recall@1 0.001. Results are averaged over subjects 1, 2, 5, and 7. FS reads the full sequence of 500 parcels, VS the 66-parcel visual subset selected at the encoder output. The first row uses no encoder and ranks the betas directly. Each subsequent row adds one component to the row above.

Method	Mean rank↓		Recall@1↑	
	FS	VS	FS	VS
Betas (no encoder)	356.68	197.76	0.01	0.09
JEPA default (same-trial, $\mathcal{L}_{\text{pred}}$ only)	392.51	362.69	0.01	0.01
+ cross-trial pairing	237.40	116.27	0.02	0.13
+ context loss $\mathcal{L}_{\text{ctx}}$	33.28	18.44	**0.39**	0.61
+ reliability weighting	**31.48**	**16.53**	**0.39**	**0.63**

every metric except SSIM in image reconstruction, which fails to separate the zeroed-betas control. This is consistent with Scotti et al. (2024), who note that SSIM is unreliable and can score well even for degenerate outputs. Decoding from the full parcel sequence and from the visual token subset (Section 3.1) performs comparably on both tasks, indicating that the heads are largely insensitive to the additional, mostly non-visual parcels. Captioning is on par with the fully supervised CorText baseline; reconstruction trails MindEye2 slightly, plausibly because its head is pre-trained on seven further subjects. Together, these results show that a representation learned without access to the stimuli supports generative decoding at a level close to pipelines trained end-to-end against them.

Next, we turn to within-subject cross-trial retrieval (Table 3). Given one stimulus repetition, this evaluation tests for retrieval of another repetition of the same stimulus, measuring how well a representation preserves stimulus-related information while suppressing trial-specific variability. We first apply it to the raw betas, which serve as a control: they ablate our pre-training entirely and establish what the task yields with no learned representation at all. We then evaluate the brain embeddings learned by our objective, along with each component's contribution. The default JEPA baseline — which draws context and target branches from the same trial and uses only the prediction loss — falls below the raw betas. Both cross-trial pairing and the context loss are therefore necessary for stimulus-discriminative representations, with the context loss yielding the largest gain. Reliability weighting adds a small further improvement, visible mainly in mean rank, with Recall@1 essentially unchanged. Finally, restricting retrieval to the visual cortex token subset substantially improves performance over the full token sequence, raising Recall@1 from 0.39 to 0.63 and reducing mean rank from 31.48 to 16.53, indicating that the remaining parcels contribute variability that is not stimulus-locked.

The same components that improve within-subject retrieval also facilitate cross-subject alignment (Table 4). Here, we map an averaged response from one subject into another subject's representation space using Eq. 1, and retrieve the matching stimulus. We first evaluate a control operating on dimensionality-reduced betas, since fitting the unsupervised transformation in the full beta space is infeasible; it scores close to chance. Our learned representations, by contrast, support accurate cross-subject retrieval. As in the within-subject setting, cross-trial pairing and the context loss account for most of the improvement, with reliability weighting adding a small further gain. Whitening the representations after pre-training improves recall and mean rank further still, to a Recall@1 of 0.59 at a mean rank of 7.45. Because a separate rotation is fitted for every token, aligning the full 500-parcel sequence is considerably more expensive than aligning the 66-token visual subset, and we therefore fit it only for the complete pipeline; alignment remains possible there, reaching Recall@1 of 0.38 at a mean rank of 17.45. The full sequence retrieves less accurately than the visual subset by a similar margin as within subjects (Table 3), so the drop reflects the lower stimulus specificity of the mostly non-visual parcels rather than any added difficulty in aligning them. Retrieval without any learned transformation, finally, performs close to chance: independently trained subject representations are not aligned by default, yet their geometry is compatible enough that orthogonal maps recover the correspondence.

Retrieval shows that the aligned spaces correspond, but the question alignment is ultimately meant to answer is whether a decoding head can read them. We therefore transfer a head trained on one

Table 4: Pre-training and alignment ablation, measured by cross-subject retrieval. For each pair of subjects, an orthogonal map is fitted per parcel token, from each subject's own training embeddings alone and with no shared stimuli or labels; a subject's repetition-averaged response is then mapped into the other's frame and ranked by cosine similarity against that subject's embeddings over the full 1000-stimulus test set. Chance is mean rank 500.5 and Recall@1 0.001. Results are averaged over the 12 ordered pairs of subjects 1, 2, 5, and 7. The first row uses no encoder and aligns dimensionality-reduced betas directly. Each subsequent row adds one component to the row above, except the last, which leaves pre-training unchanged and whitens the embeddings after training as part of the alignment procedure. All rows use the 66-parcel visual subset selected at the encoder output.

Method	Mean rank↓	Recall@1↑
Betas (no encoder) + PCA	429.07	0.00
JEPA default (same-trial, $\mathcal{L}_{\text{pred}}$ only)	452.40	0.00
+ cross-trial pairing	367.17	0.05
+ context loss $\mathcal{L}_{\text{ctx}}$	15.00	0.45
+ reliability weighting	13.11	0.49
+ whitening (post-training)	**7.45**	**0.59**

subject to another without any refinement, and evaluate it on both decoding tasks (Tables 1 and 2; results for all subject pairs in Appendix B and example reconstructions in Appendix C). As a control (*visual-readout w/o alignment*), we transfer the head directly between subject encoders, ablating the alignment transformation itself; performance collapses, confirming that the independently learned subject spaces are not directly interchangeable. Mapping the other subject's representations into the space on which the head was trained substantially recovers decoding performance.

Within participants, the visual cortex token subset and the full sequence perform equivalently on both tasks. Across participants, they diverge, and not in the same direction: the visual subset transfers better for captioning (0.453 vs. 0.394 CIDEr), whereas the full sequence transfers slightly better for reconstruction on the high-level metrics ($+1.2$ CLIP, $+1.1$ Inception). Since both decoders read the same brain embeddings through the same alignment frames, the difference must lie in how they consume the token sequence. We attribute it to the pooling stage: the reconstruction head pools its input attentively and can learn which parcels to weight, whereas the captioning projector applies fixed weights to every token. Poorly aligned non-visual parcels would then be discounted in the first case and injected as noise in the second.

5 Conclusion

Latent prediction offers a common representational basis for expressive within-subject decoding and for generalization across subjects. We trained one JEPA encoder per NSD participant, predicting the latent representation of masked cortical parcels from a different repetition of the same stimulus. The encoder never sees the stimulus images or any representation derived from them. Kept frozen, the encoder supports both brain captioning and image reconstruction at a level close to CorText and MindEye2, whose brain encoder is trained end-to-end on text or image targets.

Encoders trained independently for different subjects can also be aligned by orthogonal maps estimated without paired data. After alignment, a decoding head trained on one subject can be applied unchanged to another, retaining most of its within-subject quality on captioning and reconstruction alike; without alignment, decoding collapses. Marcos-Manchón et al. (2026) showed that compact embeddings built for retrieval admit such maps. Our results extend this to representations rich enough to drive generative decoders — which is what makes the alignment useful rather than merely measurable.

Because neither pre-training nor alignment uses the content of the stimuli, and alignment needs no stimuli shared between subjects, the approach can in principle combine data from subjects who saw different stimuli. Future work will test whether pooling data in this way improves the learned representations, whether alignment holds across datasets with different stimulus distributions, whether the objective extends to data without repeated presentations, and whether the encoders can be adapted to subjects unseen during pre-training.

AI USE STATEMENT

In this work, we used generative AI tools to assist with writing. Specifically, we first drafted all sections, then used AI to refine and polish the draft. We also used AI to source relevant related work. We manually reviewed all references sourced this way. Finally, we also used AI to assist with debugging and reviewing code. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, clean and reformat datasets, support qualitative and thematic data analysis, or interpret results. The tasks of generating synthetic data sets, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, and assisting with translation are not applicable to this work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses the Natural Scenes Dataset, a publicly released fMRI dataset collected with participant consent (Allen et al., 2022). We collected no new data from human participants and release no new data. Furthermore, this work builds on previous findings in brain decoding, and we do not believe any of its contributions raise specific ethical concerns.

REPRODUCIBILITY STATEMENT

All data we used is publicly available, and we explain our data preparation, pre-training setup, and downstream training in Section 3. Appendix A lists specific training parameters. Upon publication, we will make our pre-training code available to enable reproducibility.

REFERENCES

Emily J. Allen, Yihan Wu, J. Benjamin Hutchinson, Thomas Naselaris, and Kendrick N. Kay. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, 2022. doi: 10.1038/s41593-021-00962-x.

David Alvarez-Melis and Tommi Jaakkola. Gromov-Wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1881–1890, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. URL https://aclanthology.org/D18-1214/.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael G. Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629, 2023.

Mido Assran et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael G. Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

T. Bazeille, H. Richard, H. Janati, and B. Thirion. Local optimal transport for functional brain template estimation. In *Information Processing in Medical Imaging*, pp. 237–248, Cham, 2019. Springer International Publishing. ISBN 978-3-030-20351-1.

V. Bosch, D. Anthes, A. Doerig, S. Thorat, P. König, and T. C. Kietzmann. Brain-language fusion enables interactive neural readout and in-silico experimentation. *arXiv preprint arXiv:2509.23941*, 2025. URL https://arxiv.org/abs/2509.23941.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

Jiaxuan Chen, Yu Qi, Yueming Wang, and Gang Pan. Mindgpt: Interpreting what you see with non-invasive brain recordings. *arXiv preprint arXiv:2309.15729*, 2023a.

Po-Hsuan Chen, Xia Zhu, Hejia Zhang, Javier S. Turek, Janice Chen, Theodore L. Willke, Uri Hasson, and Peter J. Ramadge. A convolutional autoencoder for multi-subject fmri data aggregation, 2016. URL https://arxiv.org/abs/1608.04846.

Po-Hsuan (Cameron) Chen, Janice Chen, Yaara Yeshurun, Uri Hasson, James Haxby, and Peter J Ramadge. A reduced-dimension fmri shared response model. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/b3967a0e938dc2a6340e258630febd5a-Paper.pdf.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 1597–1607, 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Zijiao Chen, Jiaxin Qing, Tiange Xiang, Wan Lin Yue, and Juan Helen Zhou. Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22710–22720, June 2023b.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/forum?id=H196sainb.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Zijian Dong, Ruilin Li, Yilei Wu, Thuan Tinh Nguyen, Joanna Su Xian Chong, Fang Ji, Nathanael Ren Jie Tong, Christopher Li Hsian Chen Chen, and Juan Helen Zhou. Brain-jepa: Brain dynamics foundation model with gradient positioning and spatiotemporal masking. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Matteo Ferrante, Furkan Ozcelik, Tommaso Boccato, Rufin VanRullen, and Nicola Toschi. Brain captioning: Decoding human brain activity into images and text. *arXiv preprint arXiv:2305.11560*, 2023. URL https://arxiv.org/abs/2305.11560.

Matthew F. Glasser, Timothy S. Coalson, Emma C. Robinson, Carl D. Hacker, John Harwell, Essam Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F. Beckmann, Mark Jenkinson, Stephen M. Smith, and David C. Van Essen. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, Aug 2016. doi: 10.1038/nature18933.

Jean-Bastien Grill, Félicien Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Korukonda, S. M. Ali Calian, Michaël Caucheteux, Josip Djolonga, Florian Ströbl, Aaron van den Oord, Sylvain Gelly, Hugo Larochelle, and Remi Munos. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 21271–21284, 2020.

James V Haxby, J Swaroop Guntupalli, Andrew C Connolly, Yaroslav O Halchenko, Bryan R Conroy, M Ida Gobbini, Michael Hanke, and Peter J Ramadge. A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, 72(2):404–416, 2011. doi: 10.1016/j.neuron.2011.08.026.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, June 2022.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Jessie Huang, Erica L. Busch, Tom Wallenstein, Michal Gerasimiuk, Andrew Benz, Guillaume Lajoie, Guy Wolf, Nicholas B. Turk-Browne, and Smita Krishnaswamy. Learning shared neural manifolds from multi-subject fmri data. 2021. URL https://arxiv.org/abs/2201.00622.

Wanqiu Huang. Brainchat: Decoding semantic information from fmri using vision-language pretrained models. *arXiv preprint arXiv:2406.07584*, 2024. URL https://arxiv.org/abs/2406.07584.

Rishi Jha, Collin Zhang, Vitaly Shmatikov, and John X. Morris. Harnessing the universal geometry of embeddings. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Peter Yongho Kim, Junbeom Kwon, Sunghwan Joo, Sangyoon Bae, Donggyu Lee, Yoonho Jung, Shinjae Yoo, Jiook Cha, and Taesup Moon. Swift: Swin 4d fmri transformer. In *Advances in Neural Information Processing Systems*, volume 36, pp. 42015–42037. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/8313b1920ee9c78d846c5798c1ce48be-Abstract-Conference.html.

Yann LeCun. A path towards autonomous machine intelligence. *OpenReview [Position paper]*, 0.9.2, 2022. URL https://openreview.net/pdf?id=BZ5a1r-kVsf.

Tsung-Yi Lin et al. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014*, pp. 740–755. Springer, 2014.

Etai Littwin, Vimal Thilak, Anand Gopalakrishnan, Shuangfei Zhai, Joshua M Susskind, and Ronan Collobert. How jepa avoids noisy features: The implicit bias of deep linear self distillation networks. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/a600f0f740605205133553cb74a1c107-Abstract-Conference.html.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26296–26306, June 2024.

Weijian Mai and Zhiping Zhang. Unibrain: Unify image reconstruction and captioning all in one diffusion model from human brain activity. *arXiv preprint arXiv:2308.07428*, 2023. URL https://arxiv.org/abs/2308.07428.

Pablo Marcos-Manchón, Rishi Jha, and Lluís Fuentemilla. Platonic representations in the human brain: Unsupervised recovery of universal geometry. *arXiv preprint arXiv:2605.20496*, 2026.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mahmoud Assran, Koustuv Sinha, Michael Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-jepa 2.1: Unlocking dense features in video self-supervised learning. *arXiv preprint arXiv:2603.14482*, 2026.

Josue Ortega Caro et al. Brainlm: A foundation model for brain activity recordings. *bioRxiv*, 2023. doi: 10.1101/2023.09.12.557460. URL https://www.biorxiv.org/content/10.1101/2023.09.12.557460v1.

Furkan Ozcelik and Rufin van Rullen. Brain-diffuser: Natural scene reconstruction from fmri signals using generative latent diffusion. *ArXiv*, abs/2303.05334, 2023. URL https://api.semanticscholar.org/CorpusID:257427435.

Angeliki Papathanasiou, Jascha Achterberg, and Thomas E. Nichols. Task-guided cross-subject latent alignment: a multi-encoder-decoder vae. *arXiv preprint arXiv:2606.15989*, 2026. URL https://arxiv.org/abs/2606.15989.

Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2536–2544, 2016. doi: 10.1109/CVPR.2016.278.

Weikang Qiu, Zheng Huang, Haoyu Hu, Aosong Feng, Yujun Yan, and Zhitao Ying. MindLLM: A subject-agnostic and versatile model for fMRI-to-text decoding. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 50572–50593. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/qiu25e.html.

Alexander Schaefer, Ru Kong, Evan M Gordon, Timothy O Laumann, Xi-Nian Zuo, Avram J Holmes, Simon B Eickhoff, and B T Thomas Yeo. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri. *Cerebral Cortex*, 28(9):3095–3114, 2018. doi: 10.1093/cercor/bhx179.

Paul Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, ethan cohen, Aidan Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth A. Norman, and Tanishq Abraham. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 25330–25355. Curran Associates, Inc., 2023.

Paul Steven Scotti, Mihir Tripathy, Cesar Torrico, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A. Norman, and Tanishq Mathew Abraham. MindEye2: Shared-subject models enable fMRI-to-image with 1 hour of data. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 44038–44059. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/scotti24a.html.

Yu Takagi and Shinji Nishimoto. 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*, pp. 14453–14463, 2023.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning (ICML '08)*, pp. 1096–1103, New York, NY, USA, 2008. ACM.

Cheng Wang, Yu Jiang, Zhihao Peng, Chenxin Li, Changbae Bang, Lin Zhao, Wanyi Fu, Jinglei Lv, Jorge Sepulcre, Carl Yang, Lifang He, Tianming Liu, Xue-Jun Kong, Quanzheng Li, Daniel S. Barron, Anqi Qiu, Randy Hirschtick, Byung-Hoon Kim, Hongbin Han, Xiang Li, and Yixuan Yuan. Scaling vision transformers for functional mri with flat maps. *arXiv preprint arXiv:2510.13768*, 2025.

Cheng Wang, Yu Jiang, Zhihao Peng, Chenxin Li, Chang-bae Bang, Lin Zhao, Wanyi Fu, Jinglei Lv, Jorge Sepulcre, Carl Yang, et al. Towards a general-purpose foundation model for functional mri analysis. *Nature Biomedical Engineering*, pp. 1–12, 2026a.

Mo Wang, Wen-Hao Ye, Junfeng Xia, Minghao Xu, Hongkai Wen, and Quanying Liu. Omni-fmri: A universal atlas-free fmri foundation model. *arXiv preprint arXiv:2601.23090*, 2026b.

Navve Wasserman, Roman Beliy, Roy Urbach, and Michal Irani. Functional brain-to-brain transformation without shared stimuli. *NeuroImage*, 327:121741, 2026. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2026.121741. URL https://www.sciencedirect.com/science/article/pii/S1053811926000595.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9653–9663, 2022.

Yanwu Yang, Chenfei Ye, et al. Brainmass: Advancing brain network analysis for diagnosis with large-scale self-supervised learning. *IEEE Transactions on Medical Imaging*, 43(11):4004–4016, 2024. doi: 10.1109/TMI.2024.3415124.

A TRAINING PARAMETERS

We use the same settings for all participants across pre-training, brain captioning, image reconstruction, and alignment. During pre-training, we train on paired trials, with the context and target encoders receiving complementary single trials. During downstream training, alignment, and evaluation, we average all three repetitions per stimulus before encoding.

Table 5: Pre-training hyperparameters for all four participants.

Parameter	Value
Architecture	
Encoder transformer depth	2
Encoder attention heads	8
Parcellation	Schaefer-500
Encoder embedding dimension	512
Embedding MLP hidden dimension	256
Predictor depth	4
Predictor embedding dimension	384
Predictor attention heads	8
Masking	
Context blocks	1
Target blocks	4
Context mask scale	0.5 – 0.7
Target mask scale	0.2 – 0.3
Optimization	
Optimizer	AdamW
Learning rate schedule	warmup + cosine
Initial learning rate	5×10^{-5}
Peak learning rate	1×10^{-3}
Final learning rate	1×10^{-6}
Warmup epochs	40
Weight decay	$0.04 \rightarrow 0.4$ (cosine)
Batch size	512
Epochs	300
EMA momentum	$0.996 \rightarrow 1.0$

Table 6: Image reconstruction optimization parameters used within the MindEye2 pipeline.

Parameter	Value
Optimizer	AdamW
Learning rate schedule	cosine
Peak learning rate	3×10^{-4}
Minimum learning rate	0
Weight decay	0.01
Batch size	64
Epochs	450

Table 7: Brain captioning optimization parameters used within the CorText pipeline.

Parameter	Value
Optimizer	AdamW (8-bit)
Learning rate schedule	cosine
Stage-1 learning rate	2×10^{-3}
Stage-2 learning rate	2×10^{-5}
Stage-1 weight decay	0.05
Stage-2 weight penalty	0.001
Stage-1 epochs	20
Stage-2 epochs	5
Batch size	32
LoRA rank	16
LoRA α	16
LoRA dropout	0.05
LoRA target modules	`q,k,v,o_proj`
Quantization	4-bit

Table 8: Cross-subject alignment hyperparameters. The map is fitted for each ordered pair of participants from their own training embeddings alone, with no shared stimuli, no paired samples, and no labels using the framework introduced by Marcos-Manchón et al. (2026).

Parameter	Value
Initialization	
k-means clusters	40
Anchor restarts	100
Random seeds	5
Refinement	
ICP iterations	100
ICP step size	$0.5 \rightarrow 0.02$
ICP neighbours	20

B PER-PARTICIPANT RESULTS

Table 9: Per-subject caption generation scores from brain activity, within-subject. Higher is better for all metrics; the best result in each column and subject is shown in bold. The final block averages over subjects and corresponds to the within-subject rows of Table 1.

		BLEU				METEOR	ROUGE-L	CIDEr	CLIP-S	RefCLIP-S
Subject	Method	1	2	3	4					
1	Full sequence	**0.630**	**0.443**	**0.310**	**0.217**	**0.201**	**0.465**	**0.689**	**0.642**	**0.711**
	Visual readout	0.622	0.436	0.301	0.210	0.199	0.461	0.681	0.639	0.709
	Trial permutation	0.388	0.171	0.071	0.033	0.088	0.291	0.050	0.391	0.471
	Zeroed betas	0.181	0.079	0.026	0.006	0.094	0.191	0.001	0.351	0.414
2	Full sequence	**0.629**	**0.445**	**0.311**	**0.218**	**0.205**	**0.469**	**0.683**	**0.634**	**0.704**
	Visual readout	0.614	0.431	0.294	0.203	0.197	0.456	0.647	0.629	0.699
	Trial permutation	0.390	0.178	0.083	0.041	0.091	0.297	0.065	0.396	0.476
	Zeroed betas	0.197	0.071	0.029	0.013	0.098	0.185	0.001	0.375	0.458
5	Full sequence	0.631	0.439	0.297	0.204	0.205	0.462	0.702	0.652	0.720
	Visual readout	**0.653**	**0.472**	**0.332**	**0.237**	**0.215**	**0.487**	**0.771**	**0.656**	**0.725**
	Trial permutation	0.392	0.178	0.076	0.035	0.091	0.296	0.057	0.395	0.475
	Zeroed betas	0.185	0.076	0.025	0.010	0.090	0.192	0.002	0.363	0.432
7	Full sequence	**0.612**	**0.426**	**0.295**	**0.210**	**0.195**	**0.459**	**0.617**	**0.611**	**0.683**
	Visual readout	0.604	0.418	0.287	0.200	0.190	0.450	0.598	0.608	0.680
	Trial permutation	0.394	0.179	0.084	0.045	0.091	0.298	0.070	0.393	0.473
	Zeroed betas	0.196	0.084	0.032	0.011	0.096	0.204	0.001	0.367	0.407
Mean	Full sequence	**0.625**	0.438	0.303	0.212	**0.201**	**0.464**	0.673	**0.635**	**0.705**
	Visual readout	0.623	**0.439**	**0.304**	**0.212**	0.200	0.463	**0.674**	0.633	0.703
	Trial permutation	0.391	0.176	0.078	0.039	0.090	0.296	0.061	0.394	0.474
	Zeroed betas	0.190	0.078	0.028	0.010	0.094	0.193	0.001	0.364	0.428

Table 10: Per-pair caption generation from brain activity, cross-subject. A head trained on one subject is applied to the aligned representations of another. Higher is better for all metrics; the best result in each column and pair is shown in bold. The final block averages over the 12 ordered pairs and corresponds to the cross-subject rows of Table 1.

		BLEU				METEOR	ROUGE-L	CIDEr	CLIP-S	RefCLIP-S
Pair	Method	1	2	3	4					
$2 \to 1$	Full sequence	0.526	0.337	0.208	0.133	0.160	0.403	0.392	0.573	0.644
	Visual readout	**0.557**	**0.358**	**0.228**	**0.153**	**0.169**	**0.421**	**0.468**	**0.583**	**0.656**
	w/o alignment	0.002	0.000	0.000	0.000	0.001	0.002	0.000	0.532	0.591
$5 \to 1$	Full sequence	0.574	0.383	**0.252**	**0.172**	0.180	0.432	**0.543**	0.605	0.675
	Visual readout	**0.579**	**0.383**	0.252	0.169	**0.180**	**0.434**	0.542	**0.608**	**0.679**
	w/o alignment	0.001	0.000	0.000	0.000	0.000	0.001	0.000	0.534	0.593
$7 \to 1$	Full sequence	0.449	0.268	0.155	0.095	0.145	0.372	0.301	0.549	0.622
	Visual readout	**0.518**	**0.330**	**0.210**	**0.139**	**0.158**	**0.399**	**0.380**	**0.552**	**0.626**
	w/o alignment	0.013	0.000	0.000	0.000	0.007	0.017	0.000	0.509	0.569
$1 \to 2$	Full sequence	0.471	0.302	0.189	0.122	0.169	0.401	0.429	0.583	0.656
	Visual readout	**0.582**	**0.397**	**0.268**	**0.183**	**0.181**	**0.438**	**0.565**	**0.606**	**0.678**
	w/o alignment	0.049	0.009	0.000	0.000	0.027	0.062	0.000	0.473	0.542
$5 \to 2$	Full sequence	0.569	0.380	0.250	0.172	0.181	0.433	0.548	**0.607**	**0.679**
	Visual readout	**0.591**	**0.408**	**0.280**	**0.195**	**0.186**	**0.449**	**0.570**	0.604	0.675
	w/o alignment	0.059	0.014	0.000	0.000	0.030	0.078	0.000	0.472	0.539
$7 \to 2$	Full sequence	0.485	0.300	0.184	0.119	0.149	0.387	0.345	0.542	0.616
	Visual readout	**0.522**	**0.334**	**0.210**	**0.136**	**0.158**	**0.401**	**0.396**	**0.564**	**0.637**
	w/o alignment	0.058	0.012	0.000	0.000	0.031	0.075	0.000	0.487	0.551
$1 \to 5$	Full sequence	0.473	0.298	0.181	0.112	0.157	0.387	0.373	0.566	0.637
	Visual readout	**0.544**	**0.356**	**0.229**	**0.150**	**0.167**	**0.412**	**0.438**	**0.583**	**0.656**
	w/o alignment	0.052	0.011	0.002	0.000	0.030	0.070	0.000	0.482	0.550
$2 \to 5$	Full sequence	0.417	0.255	0.147	0.088	0.145	0.367	0.295	**0.555**	0.625
	Visual readout	**0.477**	**0.290**	**0.171**	**0.103**	**0.149**	**0.378**	**0.312**	0.554	**0.626**
	w/o alignment	0.049	0.009	0.000	0.000	0.026	0.065	0.000	0.493	0.556
$7 \to 5$	Full sequence	0.405	0.237	0.138	0.085	0.142	0.357	0.281	0.530	0.599
	Visual readout	**0.507**	**0.317**	**0.195**	**0.125**	**0.153**	**0.394**	**0.317**	**0.543**	**0.615**
	w/o alignment	0.052	0.009	0.000	0.000	0.026	0.071	0.000	0.474	0.542
$1 \to 7$	Full sequence	0.535	0.332	0.203	0.126	0.155	0.396	0.346	0.555	0.628
	Visual readout	**0.566**	**0.378**	**0.250**	**0.168**	**0.174**	**0.423**	**0.475**	**0.583**	**0.656**
	w/o alignment	0.048	0.000	0.000	0.000	0.010	0.045	0.002	0.528	0.597
$2 \to 7$	Full sequence	0.558	0.364	0.236	0.154	0.167	0.412	0.456	0.579	0.651
	Visual readout	**0.571**	**0.383**	**0.254**	**0.174**	**0.172**	**0.422**	**0.485**	**0.581**	**0.653**
	w/o alignment	0.038	0.005	0.000	0.000	0.008	0.032	0.001	0.528	0.594
$5 \to 7$	Full sequence	0.536	0.347	0.221	0.146	0.163	0.408	0.412	0.571	0.643
	Visual readout	**0.572**	**0.376**	**0.247**	**0.165**	**0.175**	**0.426**	**0.483**	**0.584**	**0.657**
	w/o alignment	0.029	0.004	0.000	0.000	0.007	0.027	0.001	0.529	0.599
Mean	Full sequence	0.500	0.317	0.197	0.127	0.159	0.396	0.393	0.568	0.639
	Visual readout	**0.549**	**0.359**	**0.233**	**0.155**	**0.168**	**0.416**	**0.453**	**0.579**	**0.651**
	w/o alignment	0.038	0.006	0.000	0.000	0.017	0.046	0.000	0.503	0.569

Table 11: Per-subject image reconstruction and retrieval, within-subject. Arrows indicate the direction of improvement; the best result in each column and subject is shown in bold. The final block averages over subjects and corresponds to the within-subject rows of Table 2. Brain retrieval is undefined for the zeroed-betas control and is omitted.

		Low-Level				High-Level				Retrieval	
Subject	Method	PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	Eff↓	SwAV↓	Image↑	Brain↑
1	Full sequence	0.309	**0.369**	**97.3**	99.2	95.3	93.7	0.631	0.352	99.9	**100.0**
	Visual readout	**0.313**	0.364	97.2	**99.2**	**95.7**	**93.8**	**0.625**	**0.349**	**99.9**	99.9
	Trial permutation	0.016	0.286	48.1	49.3	49.8	49.6	0.973	0.643	0.2	0.2
	Zeroed betas	0.018	0.255	49.8	50.9	49.7	49.7	0.977	0.700	0.4	–
2	Full sequence	0.276	0.357	96.8	99.2	95.4	93.1	0.632	**0.351**	**99.9**	**99.9**
	Visual readout	**0.277**	0.361	**97.1**	**99.3**	**95.5**	**94.0**	**0.628**	0.352	99.9	99.9
	Trial permutation	0.013	0.291	49.4	50.3	49.8	49.8	0.970	0.641	0.1	0.2
	Zeroed betas	0.067	**0.375**	49.7	49.9	50.3	49.9	0.992	0.742	0.3	–
5	Full sequence	0.233	0.345	93.6	98.3	95.5	94.0	0.626	0.357	**96.7**	97.0
	Visual readout	**0.237**	0.348	**94.9**	**98.6**	**95.7**	**94.2**	**0.625**	**0.356**	96.6	**97.2**
	Trial permutation	0.021	0.291	50.9	51.5	50.0	50.3	0.971	0.640	0.2	0.3
	Zeroed betas	-0.060	**0.428**	49.7	50.1	49.0	49.8	0.998	0.708	0.3	–
7	Full sequence	**0.230**	**0.345**	91.9	**96.8**	91.8	**90.6**	0.700	0.394	92.1	92.3
	Visual readout	0.214	0.340	**91.9**	96.7	**92.1**	90.2	**0.692**	**0.393**	**92.6**	**93.0**
	Trial permutation	0.018	0.283	49.8	50.9	51.6	49.6	0.972	0.640	0.1	0.2
	Zeroed betas	0.002	0.342	49.8	49.6	50.0	50.1	0.992	0.685	0.3	–
Mean	Full sequence	**0.262**	**0.354**	94.9	98.4	94.5	92.9	0.647	0.363	97.2	97.3
	Visual readout	0.260	0.353	**95.3**	**98.5**	**94.7**	**93.0**	**0.642**	**0.363**	**97.2**	**97.5**
	Trial permutation	0.017	0.288	49.6	50.5	50.3	49.8	0.972	0.641	0.2	0.2
	Zeroed betas	0.007	0.350	49.7	50.1	49.8	49.9	0.990	0.709	0.3	–

Table 12: Per-pair image reconstruction and retrieval, cross-subject. A head trained on one subject is applied to the aligned representations of another. Arrows indicate the direction of improvement; the best result in each column and pair is shown in bold. The final block averages over the 12 ordered pairs and corresponds to the cross-subject rows of Table 2.

		Low-Level				High-Level				Retrieval	
Pair	Method	PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	Eff↓	SwAV↓	Image↑	Brain↑
	Full sequence	0.235	0.355	**93.8**	**97.4**	**90.0**	**87.5**	**0.731**	**0.405**	**94.0**	**97.0**
$2 \to 1$	Visual readout	**0.242**	**0.382**	91.8	96.0	87.6	85.2	0.759	0.435	79.4	90.0
	w/o alignment	-0.008	0.213	55.4	59.4	54.1	54.5	0.972	0.627	0.8	0.8
	Full sequence	0.172	0.339	**87.7**	**94.7**	**90.4**	**88.0**	**0.739**	**0.419**	**57.1**	**62.4**
$5 \to 1$	Visual readout	**0.205**	**0.374**	87.1	93.2	88.8	86.6	0.753	0.436	35.0	49.4
	w/o alignment	-0.004	0.218	49.1	49.6	49.0	51.4	0.983	0.642	0.2	0.2
	Full sequence	0.165	0.319	**84.9**	**90.4**	**82.9**	**80.5**	**0.815**	**0.470**	**53.1**	**63.2**
$7 \to 1$	Visual readout	**0.188**	**0.363**	83.4	88.3	81.4	79.5	0.827	0.485	32.9	47.6
	w/o alignment	-0.005	0.182	48.8	48.4	49.9	50.9	0.982	0.643	0.4	0.5
	Full sequence	**0.240**	**0.370**	93.5	**97.5**	**90.5**	**89.3**	**0.726**	**0.410**	**95.5**	**98.0**
$1 \to 2$	Visual readout	0.232	0.357	**93.6**	96.8	89.6	86.9	0.739	0.425	94.9	97.0
	w/o alignment	-0.000	0.234	53.9	53.8	50.0	50.8	0.986	0.647	0.9	0.7
	Full sequence	**0.179**	**0.348**	**87.8**	**95.0**	**91.0**	**88.1**	**0.726**	**0.415**	**57.4**	**61.7**
$5 \to 2$	Visual readout	0.174	0.345	87.3	94.3	89.2	86.4	0.752	0.431	57.1	58.8
	w/o alignment	0.003	0.207	51.3	57.7	56.6	56.5	0.962	0.616	0.6	0.6
	Full sequence	**0.167**	0.335	84.8	**91.8**	**83.8**	**82.2**	**0.803**	**0.461**	**54.2**	**62.2**
$7 \to 2$	Visual readout	0.164	**0.338**	**85.1**	91.2	82.8	80.3	0.820	0.474	52.0	58.3
	w/o alignment	0.014	0.254	51.2	51.0	51.4	51.1	0.976	0.617	0.8	0.6
	Full sequence	0.191	0.354	**89.3**	94.6	86.5	84.4	0.770	0.436	**74.0**	**81.3**
$1 \to 5$	Visual readout	**0.212**	**0.370**	88.7	**94.7**	**87.3**	**85.0**	**0.765**	**0.435**	68.8	78.3
	w/o alignment	0.041	0.270	56.5	61.2	54.2	55.5	0.961	0.606	0.3	0.4
	Full sequence	0.170	0.352	**88.3**	**94.4**	**87.7**	**84.4**	**0.761**	**0.437**	**74.9**	**82.3**
$2 \to 5$	Visual readout	**0.189**	**0.371**	87.5	94.0	87.0	82.8	0.776	0.443	68.8	77.8
	w/o alignment	0.053	0.280	57.2	63.0	56.7	56.9	0.962	0.612	0.7	0.5
	Full sequence	0.165	0.336	**84.6**	**90.7**	**83.8**	**81.3**	**0.819**	**0.471**	**58.3**	**67.6**
$7 \to 5$	Visual readout	**0.165**	**0.363**	83.6	90.5	82.2	79.3	0.822	0.474	52.1	63.6
	w/o alignment	-0.008	0.275	47.9	47.0	50.6	50.0	0.982	0.650	0.3	0.4
	Full sequence	**0.196**	0.349	88.4	**95.0**	85.9	84.7	0.778	0.447	75.6	80.5
$1 \to 7$	Visual readout	0.189	**0.357**	**89.7**	94.6	**86.6**	**84.9**	**0.771**	**0.440**	**76.3**	**83.0**
	w/o alignment	0.045	0.240	48.7	49.5	48.1	48.9	0.992	0.672	1.0	0.9
	Full sequence	0.182	0.351	88.3	**94.6**	**86.7**	**85.0**	**0.774**	**0.442**	**73.3**	79.4
$2 \to 7$	Visual readout	**0.188**	**0.359**	**90.2**	94.3	85.2	84.3	0.781	0.446	71.9	**80.5**
	w/o alignment	0.019	0.254	49.0	51.9	49.3	49.2	0.983	0.644	0.5	0.4
	Full sequence	**0.190**	0.340	87.6	**94.7**	**88.0**	**87.0**	**0.761**	0.438	**62.1**	**70.0**
$5 \to 7$	Visual readout	0.178	**0.353**	**87.7**	93.8	87.0	86.2	0.765	**0.435**	59.4	67.4
	w/o alignment	-0.018	0.301	51.2	49.3	49.1	50.0	0.982	0.636	0.4	0.4
	Full sequence	0.188	0.346	**88.3**	**94.2**	**87.3**	**85.2**	**0.767**	**0.438**	**69.1**	**75.5**
Mean	Visual readout	**0.194**	**0.361**	88.0	93.5	86.2	84.0	0.777	0.447	62.4	71.0
	w/o alignment	0.011	0.244	51.7	53.5	51.6	52.1	0.977	0.634	0.6	0.5

C IMAGE RECONSTRUCTIONS



Figure 2: Image reconstruction from brain activity, within-subject. Reconstructions use the refined MindEye2 pipeline, whereas Table 2 reports metrics on the unrefined outputs. *Full sequence*, *Visual readout* and the control conditions are defined as in Table 1. Stimuli were drawn at random from the 1000-image test set with a fixed seed.



Figure 3: Image reconstruction from brain activity, cross-subject. Reconstructions use the refined MindEye2 pipeline, whereas Table 2 reports metrics on the unrefined outputs. The *w/o alignment* control transfers the head between encoders without the alignment transformation. The same stimuli are shown as in Figure 2.