Visual stimuli reconstruction from simultaneous fMRI-EEG signals

Daniil Dorin
MIPT
Moscow, Russia
dorin.dd.contact@gmail.com

Nikita Kiselev
MIPT
Moscow, Russia
kiselev.ns@phystech.edu

Ernest Nasyrov
MIPT
Moscow, Russia
nasyrov.rr@phystech.edu

Kirill Semkin
MIPT
Moscow, Russia
semkin.ki@phystech.edu

Abstract

How to decode human vision through neural signals has attracted a long-standing interest in neuroscience and machine learning. Modern contrastive learning and generative models improved the performance of visual decoding and reconstruction based on functional Magnetic Resonance Imaging (fMRI) and electroencephalography (EEG). However, combining these two types of information is difficult to decode visual stimuli, including due to a lack of training data. In this study, we present an end-to-end fMRI-EEG based visual reconstruction zero-shot framework, consisting of multiple tailored brain encoders and fuse module, which projects neural signals from different sources into the shared subspace as the CLIP embedding, and a two-stage multi-pipe fMRI-EEG-to-image generation strategy. In stage one, fMRI and EEG are embedded to align the high-level CLIP embedding, and then the prior diffusion model refines combined embedding into image priors. In stage two, we input this combined embedding to a pre-trained diffusion model. The experimental results indicate that our fMRI-EEG-based visual zero-shot framework achieves SOTA performance in reconstruction, highlighting the portability, low cost, and hight temporal and spatial resolution of combined fMRI-EEG, enabling a wide range of BCI applications. Our code is available in `this repository`.

1 Introduction

A key technical challenge in BCIs is to decode/reconstruct the visual world seen by humans through non-invasive brain recordings, such as fMRI, MEG or EEG. These highly dynamic brain activities reflect human perception of the visual world, which is influenced by properties of the external visual stimulus, our internal states, emotions and even personal experiences. Thus, visual decoding and reconstruction based on neural signals can uncover how the human brain processes and interprets natural visual stimulus, as well as promote non-invasive BCI applications.

Contrastive learning and generative models have greatly advanced fMRI-based visual decoding in both decoding tasks (e.g., image classification and retrieval) and generative tasks (e.g., image reconstruction). By combining pre-trained visual models, existing fMRI decoding models can learn highly-refined feature embeddings in limited data [1, 2]. Using these embedded fMRI features, generative models such as diffusion models can reconstruct the image one is seeing [2, 3].

Preprint. Under review.

At the same time EEG is portable, cheap, and universal, facilitating a wide range of BCI applications. EEG has higher temporal resolution and can effectively capture rapid changes in brain activity when processing complex, dynamic visual stimulus. It also suffers from low signal-to-noise ratio [4], low spatial resolution [5], and large inter-subject variability [6, 7]. Nonetheless, recent advances in multimodal alignment have made MEG/EEG visual decoding possible, with the performance is still inferior to fMRI [8–10]. Song et al. [9] used an EEG encoder based on ShallowNet [11] and performed representation alignment through contrastive learning, achieving excellent decoding performance on the THING-EEG dataset [12]. Li and Wei et al. [13] proposed a tailed EEG encoder, called the Adaptive Thinking Mapper (ATM), using the latest achievements of transformer architectures and achieving state-of-the-art image reconstruction quality.

Separate processing of EEG/fMRI was the only option until appearance of the recent datasets such as [14]. Now it is possible to fuse both signals and try to increase performance of the image reconstruction algorithms. To do so, we have developed a visual decoding framework based on both fMRI and EEG, including separate state-of-the-art encoders, novel fuse module, and a two-stage image generation strategy. Our contributions can be summarized as follows:

- We present brain decoding framework, which is the first work allowing zero-shot image reconstruction via *simultaneous* fMRI-EEG data. Our framework is applicable to various common fMRI and EEG encoder architectures.
- We compare different fuse module architectures to combine EEG which exhibits high temporal resolution, and fMRI that shows significantly higher spatial resolution, to achieve state-of-the-art performance in visual decoding task.
- We report a two-stage fMRI-EEG-to-image generation strategy, which separately extracts features from fMRI and EEG and refine these features with an additional lightweight prior diffusion model, enabling reliable reconstruction.

2 Method

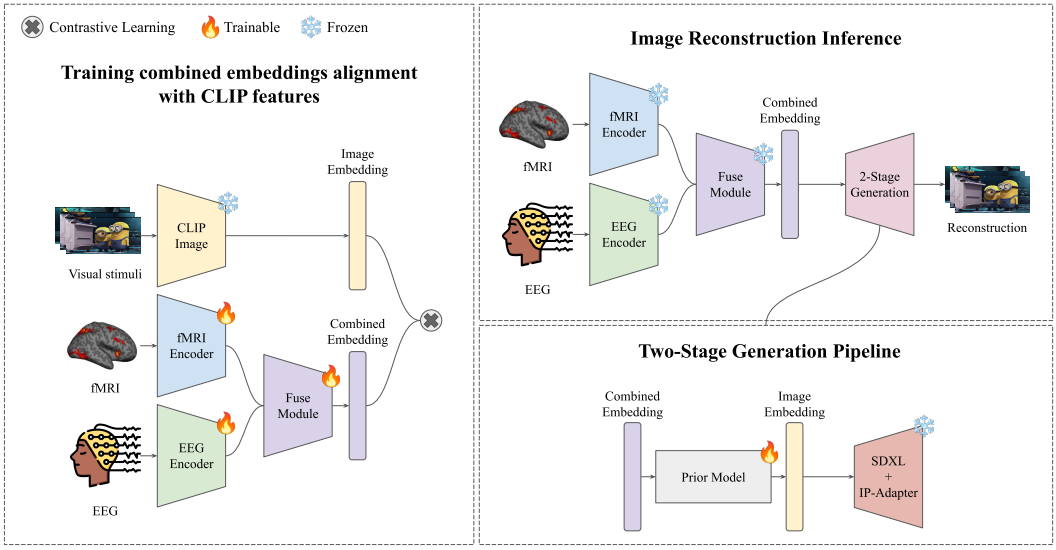

Figure 1: **fMRI-EEG-based visual decoding and generation framework.** The fMRI and EEG encoders are designed as flexible replacement components. After aligning with image features, the combined fMRI-EEG features are used to obtain reconstructed images through a two-stage generator.

To learn high-quality latent representations of both fMRI and EEG data, we use state-of-the-art encoders. EEG encoder considers the spatial position of EEG channels and the spatio-temporal properties of EEG signals. Let T represent the length of the time window of the data, C the number of EEG channels, and N the total number of data samples. Objective of EEG encoder is to derive EEG

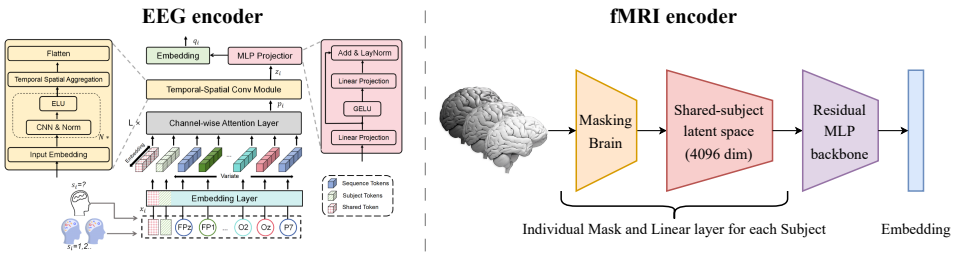

Figure 2: fMRI and EEG encoder architectures

embeddings $\mathbf{Z}_e = \mathcal{E}_e(\mathbf{E}) \in \mathbb{R}^{N \times d_e}$ from the brain activity EEG data $\mathbf{E} \in \mathbb{R}^{N \times C \times T}$, where $\mathcal{E}_e$ is the EEG encoder and d_e is the projection dimension of the embeddings. Similarly, given fMRI tensors dimensionalities X, Y, Z , for the fMRI brain activity data $\mathbf{F} \in \mathbb{R}^{N \times X \times Y \times Z}$ there is another fMRI encoder $\mathcal{E}_f$ that maps the original fMRI tensor to its latent representation $\mathbf{Z}_f = \mathcal{E}_f(\mathbf{F}) \in \mathbb{R}^{N \times d_f}$.

As separate embeddings $\mathbf{Z}_e$ and $\mathbf{Z}_f$ have been received, we pass them through the trainable **fuse module** $\mathcal{F}$, which combines high-frequency information from EEG and high spatial resolution information from fMRI, and returns the combined embeddings $\mathbf{Z}_c = \mathcal{F}(\mathbf{Z}_f, \mathbf{Z}_e) \in \mathbb{R}^{N \times d}$. Next, we will discuss how this module can be selected in order to achieve the best alignment with image embeddings.

Concurrently, we use the CLIP model [15] to extract image embeddings $\mathbf{Z}_I \in \mathbb{R}^{N \times F}$ from images. Our goal is to effectively align the fMRI-EEG representation with the image representation, as illustrated in Fig. 1. In the training phase, the fMRI encoder $\mathcal{E}_f$, EEG encoder $\mathcal{E}_e$, and fuse module $\mathcal{F}$ are trained with (fMRI, EEG, image) triplets using a contrastive learning framework. In the inference phase, the combined embeddings from the trained fused projector are used for image reconstruction.

2.1 Image Embedding

Many previous studies have explored various training strategies to train deep neural networks for image embedding, such as VGG-19 [16] and ResNet [17] trained with supervised learning, CLIP [15], DINO [18] trained with contrastive learning, and VAEs with self-supervised learning [19]. They have reported that DINO and CLIP models pre-trained using the Vision Transformer (ViT) architecture perform better in a range of downstream tasks, including image decoding and reconstruction, compared to models trained using supervised learning methods (such as VGG, ResNet) and self-supervised VAE frameworks. Thus, in this study, we use CLIP for image embedding, denoted as $\mathbf{z}_I \in \mathbb{R}^{N \times 1024}$. Before formal training, all images undergo the standard preprocessing procedure.

2.2 Contrastive training of the brain encoder

The cosine distance between embeddings is used for logits matrix calculation:

$$\mathbf{Z}_I^n = \frac{\mathbf{Z}_I}{\|\mathbf{Z}_I\|}, \quad \mathbf{Z}_c^n = \frac{\mathbf{Z}_c}{\|\mathbf{Z}_c\|}, \quad \mathbf{L} = \mathbf{Z}_I^n \left(\mathbf{Z}_c^n\right)^\intercal \cdot e^\tau$$

where τ is a trainable temperature parameter.

Contrastive Loss, similar to CLIP [20]

$$\mathcal{L} = -\frac{1}{2B} \sum_{b=0}^{B-1} \log \left(\frac{\exp(\mathbf{L}_{bb})}{\sum_{j=0}^{B-1} \exp(\mathbf{L}_{bj})} \right) + \log \left(\frac{\exp(\mathbf{L}_{bb})}{\sum_{i=0}^{B-1} \exp(\mathbf{L}_{ib})} \right)$$

We denote $\mathbf{L}_{ij}$ as an element of logits matrix $\mathbf{L}$ on the position $(i,\ j)$.

2.3 fMRI-EEG Guidance Image Generation

In this study, we present a two-stage pipeline for generating images that serve as visual stimulus for simultaneous fMRI-EEG recordings, as shown in the bottom right of Fig. 1. After fMRI encoder,

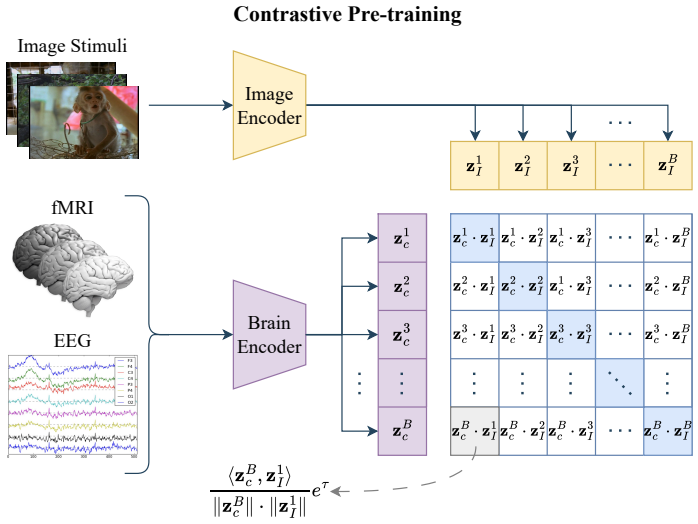

Figure 3: An illustration of contrastive learning and batching. Negative examples are obtained automatically. Cosine distances between CLIP embeddings and brain embeddings are used as logits

EEG encoder, and fuse module, we obtain combined embeddings $\mathbf{z}_c$ for each image. Now our goal is to use these combined embeddings to generate the corresponding images.

The joint distribution of images $\mathbf{x}$, combined embeddings $\mathbf{z}_c$, and image embeddings $\mathbf{z}_I$ can be expressed as $p(\mathbf{x}, \mathbf{z}_c, \mathbf{z}_I) = p(\mathbf{z}_I|\mathbf{z}_c)p(\mathbf{x}|\mathbf{z}_I)$, corresponding to the prior diffusion and CLIP-guided generation, respectively.

In **Stage I**, we first focus on the prior diffusion stage. Inspired by DALL-E 2 [21] and Mind's Eyes [2], we train a diffusion model conditioned on the combined embeddings $\mathbf{z}_c$ to learn the distribution of CLIP embeddings $p(\mathbf{z}_I|\mathbf{z}_c)$. In this stage, we construct a lightweight U-Net: $\boldsymbol{\epsilon}_{\text{prior}}(\mathbf{z}_I^t, t, \mathbf{z}_c)$, where $\mathbf{z}_I^t$ represents the noise CLIP embeddings at diffusion time step t. We train the prior diffusion model using combined and CLIP embeddings. Through the diffusion model, we can generate corresponding CLIP embeddings $\mathbf{z}_I$ from combined embeddings as a prior for stage II.

In **Stage II**, we employ the pre-trained SDXL [22] and IP-Adapter [23] models to model the generator $p(\mathbf{x}|\mathbf{z}_I)$, thereby sampling image $\mathbf{x}$ according to $\mathbf{z}_I$.

3 Data

Dataset of simultaneous recordings of fMRI and EEG was first created in [14]. It includes simultaneous recordings from 22 individuals (ages 23 to 51). Such recordings include rest, the visual paradigm Inscapes, and several short video movies representing naturalistic stimuli.

Authors provide the following information about the structure of their data:

Table 1: Dataset description

Modality	Equipment	Sampling Rate	Channels
EEG	Brain Products BrainCap MR	5000 Hz	61 cortical electrodes
fMRI	Siemens 3.0 T TIM Trio	0.476 Hz	12-channel head coil

3.1 EEG Data Processing

The total number of channels, used in EEG is 61. But in reality, some recordings were made, including inly a subset of the full channel list. We found, that not less when 50 out of 61 channels were used in the recordings, but in each recording different channels were absent. But in ML pipeline data should

be always in one format, so, we should use the same channels throughout our pipeline, not depending on the data.

We found that if we take intersections of all channel presented, we would end up with an *empty* list of channels.

To overcome this difficulty, we recover absent channel's signal based on present channel signals, using unsupervised approach. The idea is simple: to approximate the missing channel we need average signals of nearest channels with some weights.

We compared 4 recovering strategy:

- **Zero**: missing values are set to be 0.
- **NN**: missing channel signal is a copy of the nearest existing channel signal.
- **KNN Euclidean**: missing channel signal is a weighted average of k nearest existing channel signals with Euclidean weights.
- **KNN Spherical**: missing channel signal is a weighted average of k nearest existing channel signals with spherical weights.

The results of our approach are presented in Fig. 4.

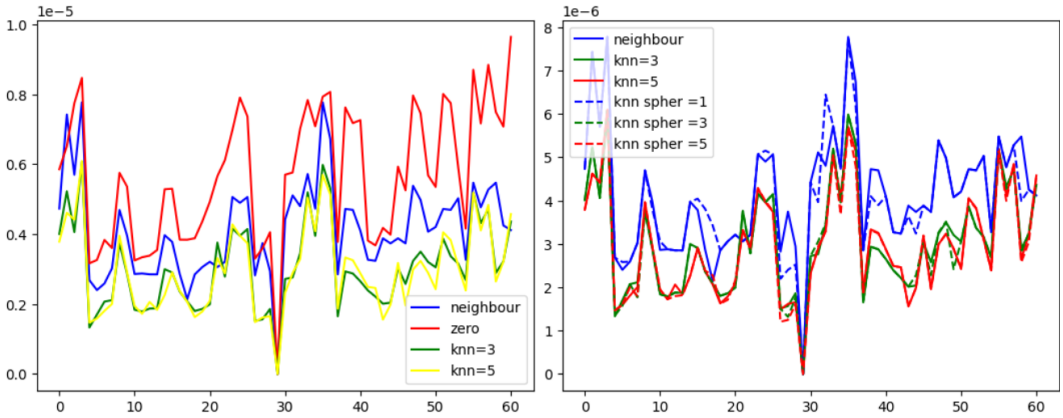

Figure 4: OX: number of channel to be recovered. OY: MSE distance from real channel signal to its recovered version. Labels denote used methods. *neighbour* stands for *NN* strategy, *knn=*K for *KNN Euclidean*, *knn spher=*K for *KNN Spherical*, *zero* for *Zero*.

We see, that *Zero* strategy is the words, having the biggest reconstruction error among all methods, and *knn*-family are the best. Practically, we do not found significant difference between sprerical and euclidean knn, so we have chosen the former for its simplicity.

4 Experiments

To verify the theoretical estimates obtained, we conduct a detailed empirical study.

4.1 Contrastive Learning results

The paper analyzes the quality of recovery, both with and without a diffusion prior. The results with a diffusion prior are shown in the Figure 6.

In the initial experiment, we were not able to achieve satisfactory results.

4.2 Diffusion Prior results

Thus, we do the diffusion prior training. It is applied in the latent space to align the obtained combined embeddings $\mathbf{z}_c$ with image embeddings $\mathbf{z}_I$.

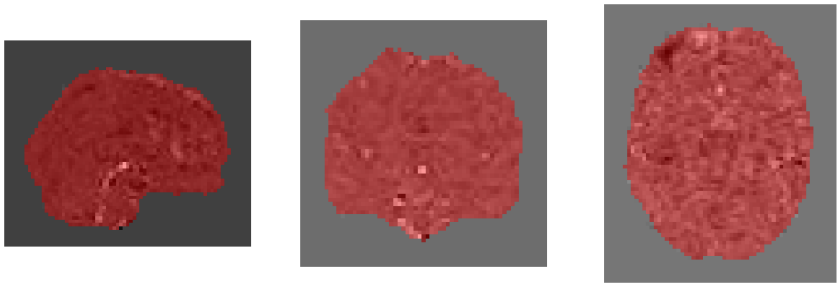

Figure 5: To reduce the dimensionality of fMRI data, brain masks are considered to exclude a background that does not carry the necessary information about the stimulus. Masks are calculated based on the analysis of the most changing voxels.

Figure 6: Examples of image reconstruction without a diffusion prior

Figure 7: Examples of image reconstruction after diffusion prior

Our results are presented in the Figure 7.

After training a diffusion prior module, we do not achieve high quality yet... Thus, we want to change our experiments setup...

4.3 Alignment metrics calculation

To assess the model quality, we utilize CLIP-Score, as we align combined embeddings with CLIP image embeddings:

$$\text{CLIP-Score}(\mathbf{z}_c, \mathbf{z}_I) = \frac{\langle \mathbf{z}_c, \mathbf{z}_I \rangle}{\|\mathbf{z}_c\| \|\mathbf{z}_I\|}.$$

This score reflects the cosine distance between obtained embeddings and initial image embeddings. We calculate this score for each triplet in the dataset, obtaining the results in the table below.

Model	dme	dmh	tp	inscapes	monkey1	monkey2	monkey5
fMRI-EEG	0.099	0.097	0.097	0.089	0.071	0.113	0.104
EEG (monkeys)	–	–	–	–	0.037	0.058	0.054

The obtained results show that training separate EEG encoder has worse results compare to the combined fMRI-EEG brain encoder.

5 Related Work

Datasets with Simultaneous fMRI-EEG Signals. The availability of high-quality datasets with fMRI and EEG signals is crucial for advancing research in brain-computer interfaces (BCIs). Several recent studies have provided valuable datasets that include both fMRI and EEG recordings, which are essential for training and validating models that decode visual stimuli from neural signals. These datasets cover a range of visual and naturalistic stimuli, including flickering checkerboards, visual paradigms, and short video movies [14, 24, 25]. Other datasets focus on naturalistic stimuli such as audiovisual films and face processing tasks [26–29]. However, only one of them [14] is convenient for our task, as it contains **simultaneous** fMRI-EEG signals from 22 individuals viewed several short video movies. In this paper, we construct training triplets (fMRI, EEG, image) using this dataset.

Image reconstruction using fMRI. Recent advances in fMRI-based visual decoding have significantly improved the performance of visual reconstruction tasks. Several studies have utilized contrastive learning and generative models to achieve high-quality image reconstruction from fMRI signals. For example, Ozcelik et al. [30] and Takagi et al. [31] used diffusion models to enhance the fidelity of generated images. Other studies have employed sophisticated architectures like SC-MBM encoders and frozen CLIP models for contrastive learning [2, 32].

Image reconstruction using EEG. Recent studies have utilized various encoder architectures and generative models to get the quality of image reconstruction comparable to best fMRI encoders. Mishra et al. [33] employed CNN for EEG feature extraction and GANs to achieve high-quality image reconstruction. Paper [34] used LSTM as EEG encoder instead. Li et al. [13] made a combined encoder out of transformer, convolutional and MLP layers. As signals from different EEG channels tend to cross-correlate [35], this architecture can utilize it for non-redundant embeddings. The authors also used diffusion model as an image generator.

Image Generation. State-of-the-art methods for image generation have played a crucial role in improving the performance of visual decoding tasks. These methods often employ sophisticated architectures, such as diffusion models and variational autoencoders [36, 37], to generate high-quality images from neural signals. For example, Xu et al. [38] and Hazami et al. [39] used diffusion models and VAEs for image generation. Other studies have employed VQGAN and StyleGAN2 for high-quality image synthesis [40, 41].

6 Conclusion

The experiments have shown that CLIP loss for combined brain encoder is decreasing. It indicates that encoder is able to learn CLIP embeddings of the images in our dataset. The diffusion prior loss is decreasing as well.

Nonetheless, after applying SDXL + IP Adapter diffusion reconstructed images does not look similar to original one. The model is able to represent general colors and shapes. However, the lack of fidelity in the reconstructed images suggests that while the model captures some of the broader characteristics of the visual stimuli, it struggles with finer details and specific features that are critical for accurate reconstruction. This discrepancy may stem from several factors.

First, the quality and diversity of the training dataset could play a significant role. If the dataset does not encompass a wide range of visual stimuli or lacks sufficient examples of certain categories, the model may not generalize well to unseen images. Future experiments should consider augmenting the dataset with a more varied selection of images to improve the model's ability to capture intricate details.

Second, the architecture of the diffusion model itself might require further tuning. While SDXL + IP Adapter shows promise, exploring alternative architectures or additional layers specifically designed to enhance detail retention could yield better results. Hyperparameter optimization could also be beneficial in striking a balance between generalization and detail preservation.

Moreover, the integration of EEG and FMRI data presents unique challenges. The temporal resolution of EEG and spatial resolution of FMRI, while complementary, may introduce noise or misalignment in the data that complicates the reconstruction process. Investigating advanced preprocessing techniques or alignment strategies could help mitigate these issues and improve the coherence between brain signals and visual outputs.

Additionally, exploring different loss functions beyond CLIP loss might provide insights into how to better capture the nuances of visual perception. Incorporating perceptual loss functions that focus on human-like perception of images could help guide the model towards producing more visually coherent outputs.

Lastly, considering the interpretability of the results is crucial. Understanding which aspects of brain activity correspond to specific features in reconstructed images can provide valuable insights into visual processing mechanisms. Employing techniques such as saliency mapping or feature visualization could help elucidate the relationship between brain signals and visual features, ultimately enhancing both model performance and our understanding of neural representation.

In conclusion, while the current results demonstrate a promising direction for reconstructing images from EEG and FMRI data, there remains significant room for improvement. By addressing data quality, model architecture, preprocessing techniques, and loss function optimization, future research can enhance the accuracy and detail of reconstructed images, paving the way for more effective applications in neuroscience and beyond.

References

[1] Yulong Liu, Yongqiang Ma, Wei Zhou, Guibo Zhu, and Nanning Zheng. Brainclip: Bridging brain and visual-linguistic representation via clip for generic natural visual stimulus decoding, 2023. URL `https://arxiv.org/abs/2302.12971`.

[2] Paul S. Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Ethan Cohen, Aidan J. Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth A. Norman, and Tanishq Mathew Abraham. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors, 2023. URL `https://arxiv.org/abs/2305.18274`.

[3] Tao Fang, Qian Zheng, and Gang Pan. Alleviating the semantic gap for generalized fmri-to-image reconstruction. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 15096–15107. Curran Associates, Inc., 2023.

[4] Alois Schlögl, Mel Slater, and Gert Pfurtscheller. Presence research and eeg. 01 2002.

[5] Ramesh Srinivasan. Methods to improve the spatial resolution of eeg.

[6] Hiroshi Morioka, Atsunori Kanemura, Jun-ichiro Hirayama, Manabu Shikauchi, Takeshi Ogawa, Shigeyuki Ikeda, Motoaki Kawanabe, and Shin Ishii. Learning a common dictionary for subject-transfer decoding with resting calibration. *NeuroImage*, 111:167–178, 2015.

[7] Vinay Jayaram, Morteza Alamgir, Yasemin Altun, Bernhard Scholkopf, and Moritz Grosse-Wentrup. Transfer learning in brain-computer interfaces. *IEEE Computational Intelligence Magazine*, 11(1):20–31, 2016.

[8] Radoslaw Martin Cichy and Dimitrios Pantazis. Multivariate pattern analysis of meg and eeg: A comparison of representational structure in time and space. *NeuroImage*, 158:441–454, September 2017. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2017.07.023. URL http://dx.doi.org/10.1016/j.neuroimage.2017.07.023.

[9] Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding natural images from eeg for object recognition, 2024. URL https://arxiv.org/abs/2308.13234.

[10] Tijl Grootswagers, Ivy Zhou, Amanda K. Robinson, Martin N. Hebart, and Thomas A. Carlson. Human eeg recordings for 1, 854 concepts presented in rapid serial visual presentation streams. *Scientific Data*, 9(1), January 2022. ISSN 2052-4463. doi: 10.1038/s41597-021-01102-7. URL http://dx.doi.org/10.1038/s41597-021-01102-7.

[11] Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human Brain Mapping*, 38(11):5391–5420, August 2017. ISSN 1097-0193. doi: 10.1002/hbm.23730. URL http://dx.doi.org/10.1002/hbm.23730.

[12] 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, December 2022. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2022.119754. URL http://dx.doi.org/10.1016/j.neuroimage.2022.119754.

[13] Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, Haoyang Qin, and Quanying Liu. Visual decoding and reconstruction via eeg embeddings with guided diffusion, 2024. URL https://arxiv.org/abs/2403.07721.

[14] Qawi K. Telesford, Eduardo Gonzalez-Moreira, Ting Xu, Yiwen Tian, Stanley J. Colcombe, Jessica Cloud, Brian E. Russ, Arnaud Falchier, Maximilian Nentwich, Jens Madsen, Lucas C. Parra, Charles E. Schroeder, Michael P. Milham, and Alexandre R. Franco. An open-access dataset of naturalistic viewing using simultaneous eeg-fmri. *Scientific Data*, 10(1), August 2023. ISSN 2052-4463. doi: 10.1038/s41597-023-02458-8. URL http://dx.doi.org/10.1038/s41597-023-02458-8.

[15] 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. URL https://arxiv.org/abs/2103.00020.

[16] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. URL https://arxiv.org/abs/1409.1556.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

[18] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021. URL https://arxiv.org/abs/2104.14294.

[19] Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. Brain decoding: toward real-time reconstruction of visual perception, 2024. URL https://arxiv.org/abs/2310.19812.

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[21] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. URL https://arxiv.org/abs/2204.06125.

[22] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. URL https://arxiv.org/abs/2307.01952.

[23] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2308.06721.

[24] Daniel G Wakeman and Richard N Henson. A multi-subject, multi-modal human neuroimaging dataset. *Scientific Data*, 2(1), January 2015. ISSN 2052-4463. doi: 10.1038/sdata.2015.1. URL http://dx.doi.org/10.1038/sdata.2015.1.

[25] Julia Berezutskaya, Mariska J. Vansteensel, Erik J. Aarnoutse, Zachary V. Freudenburg, Giovanni Piantoni, Mariana P. Branco, and Nick F. Ramsey. Open multimodal ieeg-fmri dataset from naturalistic stimulation with a short audiovisual film. *Scientific Data*, 9(1), March 2022. ISSN 2052-4463. doi: 10.1038/s41597-022-01173-0. URL http://dx.doi.org/10.1038/s41597-022-01173-0.

[26] Yameng Gu, Lucas E. Sainburg, Feng Han, and Xiao Liu. Simultaneous eeg and functional mri data during rest and sleep from humans. *Data in Brief*, 48:109059, June 2023. ISSN 2352-3409. doi: 10.1016/j.dib.2023.109059. URL http://dx.doi.org/10.1016/j.dib.2023.109059.

[27] Jonathan Gallego-Rudolf, María Corsi-Cabrera, Luis Concha, Josefina Ricardo-Garcell, and Erick Pasaye-Alcaraz. Simultaneous and independent electroencephalography and magnetic resonance imaging: A multimodal neuroimaging dataset. *Data in Brief*, 51:109661, December 2023. ISSN 2352-3409. doi: 10.1016/j.dib.2023.109661. URL http://dx.doi.org/10.1016/j.dib.2023.109661.

[28] Mohammad Momenian, Zhengwu Ma, Shuyi Wu, Chengcheng Wang, Jonathan Brennan, John Hale, Lars Meyer, and Jixing Li. Le petit prince hong kong (lpphk): Naturalistic fmri and eeg data from older cantonese speakers. April 2024. doi: 10.1101/2024.04.24.590842. URL http://dx.doi.org/10.1101/2024.04.24.590842.

[29] Umit Keles, Julien Dubois, Kevin J. M. Le, J. Michael Tyszka, David A. Kahn, Chrystal M. Reed, Jeffrey M. Chung, Adam N. Mamelak, Ralph Adolphs, and Ueli Rutishauser. Multimodal single-neuron, intracranial eeg, and fmri brain responses during movie watching in human patients. *Scientific Data*, 11(1), February 2024. ISSN 2052-4463. doi: 10.1038/s41597-024-03029-1. URL http://dx.doi.org/10.1038/s41597-024-03029-1.

[30] Furkan Ozcelik and Rufin VanRullen. Natural scene reconstruction from fmri signals using generative latent diffusion, 2023. URL https://arxiv.org/abs/2303.05334.

[31] Yu Takagi and Shinji Nishimoto. Improving visual image reconstruction from human brain activity using latent diffusion models via multiple decoded inputs, 2023. URL https://arxiv.org/abs/2306.11536.

[32] Zijiao Chen, Jiaxin Qing, and Juan Helen Zhou. Cinematic mindscapes: High-quality video reconstruction from brain activity, 2023. URL https://arxiv.org/abs/2305.11675.

[33] Rahul Mishra, Krishan Sharma, R. R. Jha, and Arnav Bhavsar. Neurogan: image reconstruction from eeg signals via an attention-based gan. *Neural Computing and Applications*, December 2022. ISSN 1433-3058. doi: 10.1007/s00521-022-08178-1. URL http://dx.doi.org/10.1007/s00521-022-08178-1.

[34] Prajwal Singh, Pankaj Pandey, Krishna Miyapuram, and Shanmuganathan Raman. Eeg2image: Image reconstruction from eeg brain signals, 2023. URL https://arxiv.org/abs/2302.10121.

[35] Ronakben Bhavsar, Yi Sun, Na Helian, Neil Davey, David Mayor, and Tony Steffert. The correlation between eeg signals as measured in different positions on scalp varying with distance. *Procedia Computer Science*, 123:92–97, 2018. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2018.01.015. URL https://www.sciencedirect.com/science/article/pii/S1877050918300164. 8th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2017 (Eighth Annual Meeting of the BICA Society), held August 1-6, 2017 in Moscow, Russia.

[36] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2021. URL https://arxiv.org/abs/2012.09841.

[37] Ali Hatamizadeh, Jiaming Song, Guilin Liu, Jan Kautz, and Arash Vahdat. Diffit: Diffusion vision transformers for image generation, 2024. URL https://arxiv.org/abs/2312.02139.

[38] Xingqian Xu, Zhangyang Wang, Eric Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model, 2024. URL https://arxiv.org/abs/2211.08332.

[39] Louay Hazami, Rayhane Mama, and Ragavan Thurairatnam. Efficient-vdvae: Less is more, 2022. URL https://arxiv.org/abs/2203.13751.

[40] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance, 2022. URL https://arxiv.org/abs/2204.08583.

[41] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan, 2020. URL https://arxiv.org/abs/1912.04958.