DECODING EEG SIGNALS OF VISUAL BRAIN REPRESENTATIONS WITH A CLIP BASED KNOWLEDGE DISTILLATION

Matteo Ferrante
University of Rome, Tor Vergata
Department of Biomedicine and Prevention
matteo.ferrante@uniroma2.it

Tommaso Boccato
University of Rome, Tor Vergata
Department of Biomedicine and Prevention

Stefano Bargione
University of Rome, Tor Vergata
Department of Biomedicine and Prevention

Nicola Toschi
University of Rome, Tor Vergata
Department of Biomedicine and Prevention
A.A. Martinos Center for Biomedical Imaging
Harvard Medical School/MGH, Boston (US)

ABSTRACT

Decoding visual representations from human brain activity has emerged as a thriving research domain, particularly in the context of brain-computer interfaces. Our study presents an innovative method that employs knowledge distillation to train an EEG classifier and reconstruct images from the ImageNet and THINGS-EEG 2 datasets using only electroencephalography (EEG) data from participants that have viewed the images themselves (i.e. ”brain decoding”). We analyzed EEG recordings from 6 participants for the ImageNet dataset and 10 for the THINGS-EEG 2 dataset, exposed to images spanning unique semantic categories. These EEG readings were converted into spectrograms, which were then used to train a convolutional neural network (CNN), integrated with a knowledge distillation procedure based on a pre-trained Contrastive Language-Image Pre-Training (CLIP)-based image classification teacher network. This strategy allowed our model to attain a top-5 accuracy of 80%, significantly outperforming a standard CNN and various RNN-based benchmarks. Additionally, we incorporated an image reconstruction mechanism based on pre-trained latent diffusion models, which allowed us to generate an estimate of the images that had elicited EEG activity. Therefore, our architecture not only decodes images from neural activity but also offers a credible image reconstruction from EEG only, paving the way for, e.g., swift, individualized feedback experiments.

1 INTRODUCTION

Electroencephalography (EEG) has gained prominence in decoding visual representations from the human brain, particularly for complex visual stimuli from datasets like ImageNet. While convolutional (CNN) and recurrent neural networks (RNN) have been effective in classifying EEG signals into image categories(Deng et al., 2009), the focus of these papers has largely been on multisubject models. Our study emphasizes single-subject models to capture individual variability in visual processing, hence offering enhanced decoding detail as well as privacy. A key challenge is reconstructing visual stimuli from EEG due to its low spatial resolution. In this respect, semantic image reconstructions (Ferrante et al., 2023; Ozcelik et al., 2022; Takagi & Nishimoto, 2023; Benchetrit et al., 2023) may be more viable than pixel-level recreations. This paper builds on prior work (Kavasidis et al.; Spampinato et al.; Palazzo et al.; Singh et al.; Bai et al.; Spampinato et al., 2019), by proposing a novel pipeline for training single-subject models for brain decoding and EEG-based image reconstruction, leveraging deep learning techniques. We address issues identified in previous studies, such as inflated performance metrics due to inadequate data preprocessing (Li et al., 2018; Bharadwaj et al., 2023; Li et al., 2021), and adhere to conservative approaches recommended by

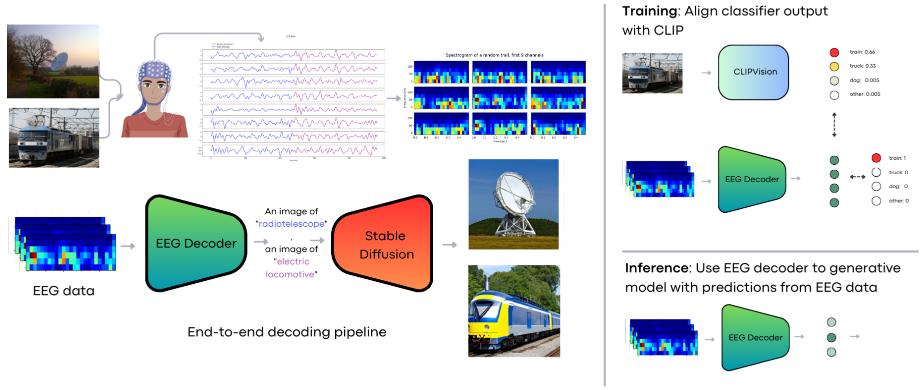
Spectrogram of a random trail, first 9 channels
EEG data
EEG Decoder
An image of "radiotelescope"
an image of "electric locomotive"
Stable Diffusion
End-to-end decoding pipeline
Training: Align classifier output with CLIP
CLIPVision
train: 0.66
truck: 0.33
dog: 0.005
other: 0.005
EEG Decoder
train: 1
truck: 0
dog: 0
other: 0
Inference: Use EEG decoder to generative model with predictions from EEG data
EEG Decoder

Figure 1: Our pipeline can be described as follows: EEG data was recorded while the subject was viewing natural images. These data are then preprocessed and converted into spectrograms, which serve as the input for our neural network. Our EEG decoder is trained using a knowledge distillation method based on the CLIP model. The outputs from the EEG decoder, which are predictions of the image that elicited the EEG data, are then combined with an image generation pipeline. This end-to-end approach allows us to reconstruct images from the neural activity data captured by the EEG.

recent re-analyses (Palazzo et al., 2020; Li et al., 2018). Our pipeline employs CLIP-based (Radford et al., 2021) knowledge distillation in a convolutional neural network, trained on time-frequency decomposition (TFD) of EEG signals, followed by generative diffusion synthesis. We therefore align neural and image representations, allowing for semantically coherent and visually similar image reconstructions. The pipeline features a flexible, independent generative component conditioned by the EEG decoder and exploits a classification bottleneck to simplify the problem and avoid direct regression of the latent diffusion model's conditioning embedding.

2 Material and Methods

This section outlines our methodology and the dataset used, sourced from ImageNet EEG (Kavasidis et al., 2017), which is publicly available. This dataset is composed of EEG recordings obtained from six participants exposed to images from 40 different ImageNet classes, each comprising 50 images. The image display protocol involved showing images sequentially for 0.5 seconds each, across 25-second intervals, always followed by a 10-second break. This resulted in showing 2,000 images over 1,400 seconds (23 minutes recording time per subject). Data collection involved a 128-channel cap with active, low-impedance electrodes (actiCAP 128Ch, Brainproducts) and Brainvision amplifiers and systems, recording EEG signals at a 1000 Hz sampling rate with 16-bit resolution. The final dataset has 11,466 sequences. This extensive experimental design, involving multi-channel EEG recordings while viewing of thousands of stimuli, provides a comprehensive dataset for training decoding models. For more details on the acquisition protocol, refer to (Kavasidis et al., 2017; Spampinato et al., 2019).

EEG data underwent a sequence of preprocessing steps after train/test (80/20%) splitting. Initially, a notch filter (49-51 Hz) reduced power line interference, and a 14-70 Hz second-order Butterworth band-pass filter isolated frequencies linked to visual attention. The signals were then standardized channel-wise. For training, we segmented the filtered EEG signals into 40 ms long windows, with 20 ms overlap. Each segment underwent time-frequency decomposition using the short-time Fourier transform (STFT or wavelet decomposition as an alternative) resulting in images that represented the time and frequency spectrum with shape (channels, time, frequency). This process yielded 1911 EEG spectrograms per subject, essential for training our CNN and classifying visual stimuli. To show generalization of our method, we also included another dataset, from the THINGS initiative collection, named THINGS-EEG2 (Gifford et al., 2022). This dataset comprises a substantial collection of EEG readings taken at high temporal precision, recording reactions to pictures of objects against a natural backdrop. It encompasses data from 10 participants, covering 82,160 instances across 16,740 different image scenarios. Image stimuli belong to the THINGS Image dataset, span-

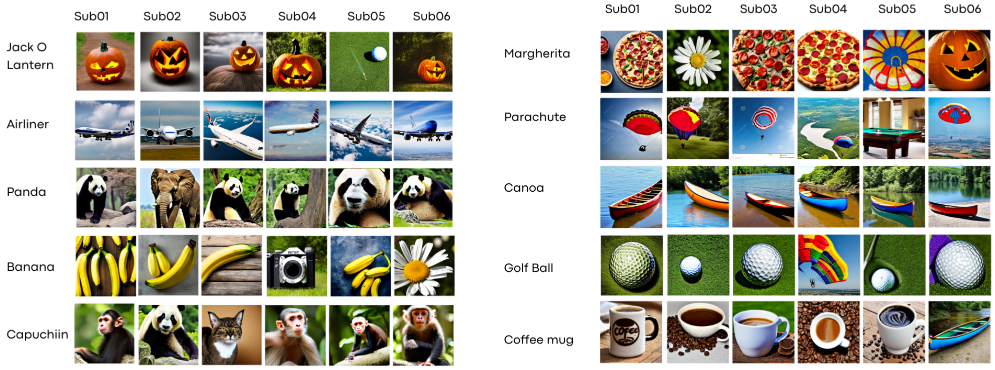

Figure 2: Reconstructed images. Left column: target classes; subsequent columns: results from individual participants starting from their EEG activity.

Method	Accuracy	Top3 Accuracy	Top5 Accuracy	F1	Kappa
LR on windowed signal	0.0205 (0.0058)	0.0636 (0.0083)	0.1092 (0.0110)	0.0156 (0.0054)	0.0009 (0.0061)
LR on PCA windowed signal	0.0175 (0.0040)	0.0536 (0.0084)	0.0961 (0.0063)	0.0097 (0.0047)	0.0020 (0.0039)
CEBRA + kNN	0.0240 (0.0050)	0.0831 (0.0116)	0.1402 (0.0136)	0.0223 (0.0061)	-0.0012 (0.0056)
LSTM	0.3605 (0.0938)	0.7376 (0.1226)	0.8868 (0.1030)	0.3392 (0.0894)	0.3437 (0.0960)
Conv1d	0.2623 (0.0511)	0.6013 (0.0826)	0.7971 (0.0851)	0.2582 (0.0520)	0.2432 (0.0524)
Knowledge distillation on eeg (img)	0.2819 (0.0836)	0.5773 (0.1379)	0.7295 (0.1339)	0.2742 (0.0794)	0.2632 (0.0857)
Knowledge distillation on wavelet	0.4060 (0.1154)	0.7490 (0.1282)	0.8787 (0.1007)	0.3889 (0.1148)	0.3905 (0.1183)
Plain CNN on spectrograms	0.2819 (0.0836)	0.5773 (0.1379)	0.7295 (0.1339)	0.2742 (0.0794)	0.2632 (0.0857)
Palazzo et al (Palazzo et al., 2020)	0.3350 (0.089)	-	-	-	-
Knowledge distillation on STFT	**0.4120 (0.1131)**	**0.7530 (0.1068)**	**0.8782 (0.0806)**	**0.4027 (0.1133)**	**0.3966 (0.1160)**
Knowledge distillation (THINGS-EEG2)	**0.58 (0.04)**	-	-	**0.52 (0.036)**	-
Plain CNN (THINGS-EEG2)	0.52 (0.03)	-	-	0.48 (0.032)	-

Table 1: Performance comparison of decoding baselines. The table presents the mean values accompanied by the standard deviation (enclosed in parentheses) for each evaluation metric across all participants. Results from (Palazzo et al., 2020) are reported from the original paper in the same setting used here. The first part of the table reports results for ImageNet-EEG dataset, while the second part report comparison between our method and plain CNN on the THINGS-EEG2 dataset.

ning across 1854 different classes. In this work, the EEG activity is recorded while 1654 categories were shown as part of the training set and the other 200 categories were shown as test set. Since the very fine granularity of concepts of this dataset makes the problem more complex, we obtained pseudo-labels for the entire dataset using a K-Means over the CLIP embeddings of all images. We used a k-Elbow approach to identify the optimal number of clusters (that turned out to be 8) and trained a K-Means to predict cluster labels to re-label this dataset. EEG data were processed as described before. Since the cluster labels cannot be used as conditioning for the generative part, we adopted a simpler approach for the second dataset, stopping our analysis with the classification part.

Knowledge distillation involves transferring knowledge from a large, pretrained teacher model to a smaller student model, enabling the latter to achieve high performance in spite of lower model capacity. (Hinton et al., 2015). For a given stimulus image x, let $f_t(x)$ be the output class probabilities from the teacher model, and $f_s(e;\theta)$ the student model's output, with θ as its parameters and e representing EEG recordings. The student model is trained by minimizing a loss function $\mathcal{L}(\theta)$, combining a cross-entropy loss $\mathcal{L}_{CE}$ with a distillation loss $\mathcal{L}_{KD}$, which measures the output difference between student and teacher models. The distillation loss $\mathcal{L}_{KD}$ includes a temperature parameter T and enhances the transfer of insights about inter-class relationships from the teacher to the student model. In our implementation, we set $\alpha = 0.5$ and $T = 1$ as optimizable hyperparameters of our procedure. Our teacher model integrates a linear classifier with CLIP (Radford et al., 2021), utilizing its image encoder (a vision transformer) to embed images into latent representations. CLIP's ability to align images and text in an embedding space is leveraged to create a classifier trained on the CLS token for image classification. The student CNN, trained on EEG data, benefits from the teacher's knowledge, focusing on neural patterns relevant to visual recognition.

Our method uses a CNN with residual connections for classifying EEG time-frequency decompositions (TFDs). It starts with convolutional layers, increasing filter numbers to extract spatial and temporal features, followed by global average pooling and fully-connected layers for classification. With our knowledge distillation method, we use an image classifier as a teacher to provide ”soft targets” for guiding our EEG model. This classifier initially predicts stimulus classes with a 99% accuracy. During training, EEG spectrograms are input into the student CNN model, while the teacher model receives CLIP image features. The aim is to align the student model’s class probability distributions with those of the teacher, enhancing stability and performance compared to direct class label training. At the inference stage, the EEG-based CNN alone predicts classes from new EEG TFDs. This knowledge distillation from the image model allows our CNN to develop robust representations for decoding visual stimuli from EEG signals. After training, our EEG model predicts ImageNet classes from new EEG TFDs. To reconstruct corresponding visual stimuli, we use the Stable Diffusion generative model (Ramesh et al., 2022). For each EEG prediction, a text prompt like ”an image of a predicted class” is created and fed into Stable Diffusion, generating images matching the predicted class. This approach allows visual stimulus reconstruction solely from neural activity. The EEG decoder determines the class, and Stable Diffusion produces an image which is semantically coherent with the class. The entire decoding pipeline is illustrated in Fig 1. This approach facilitates the synthesis of plausible image reconstructions based on the decoded semantic category from neural activity patterns. This model-centric strategy also addresses the inherent resolution constraints of EEG for high-fidelity decoding. Finally, the guided diffusion modeling ensures the generation of visualizations that are both realistic and interpretable to human observers.

In terms of baselines and comparisons, we explored various methods for EEG signal decoding, ranging from traditional machine learning to advanced neural network architectures . We started with basic logistic regression classifiers, employing techniques like standardization, averaging of raw EEG signals, PCA for component retention, and sliding window averaging. Delving into deep learning, we used the CEBRA technique (Schneider et al., 2023), which projects EEG data onto a 32-dimensional space using CLIP features to generate a nonlinear neural baseline. We also tested Recurrent Networks (LSTMs) and 1D CNN models with four layers and dropout regularization, directly processing EEG time series. Additionally, we implemented computer vision techniques using Convolutional Neural Networks (CNNs) by interpreting EEG signals as 2D images. We varied the time-frequency decomposition strategy, alternating between Short-Time Fourier Transform (STFT) and wavelet decomposition, specifically employing the Daubechies db4 wavelet (Lee et al., 2019). This methodology efficiently leveraged the structural properties of multi-channel EEG data. All neural networks had a similar parameter count (1.1-1.2 M) and were trained with the Adam optimizer, a learning rate of $3e-4$, and common specifications including early stopping, batch sizing, gradient clipping, and a max epoch limit. This diverse set of methods aimed to offer a detailed comparison between methods, underscoring the importance of spatiotemporal modeling in EEG decoding.

3 RESULTS

Our model’s effectiveness is assessed using a range of metrics: top-5, top-3, top-1 accuracy, F1 score, and normalized kappa score. Our approach, which combines a CNN on time-frequency decompositions (TFD) with CLIP-based knowledge distillation, outperforms all baselines as well as the same network without distillation, as detailed Table 2. Our results reveal several trends: classical machine learning techniques using averaged or PCA-reduced EEG data show near chance-level accuracy, highlighting the limitations of hand-engineered features in decoding complex visual stimuli. In contrast, deep learning models, particularly those handling spatiotemporal EEG TFDs, display markedly better accuracy. CNNs processing raw EEG time series or 2D multi-channel EEG representations, especially those utilizing TFDs from wavelet-transformed or spectrogram images, achieve over 85% top-5 accuracy, demonstrating the efficacy of computer vision techniques in EEG signal processing. Deep learning models considerably outperform classical methods in top-3 and top-5 accuracy metrics. Top CNNs achieve over 75% top-3 accuracy, showing that the true label often falls within the top three predictions. This indicates the proficiency of 2D convolutions in extracting semantic categories from EEG, with a noticeable performance disparity compared to LSTM networks. Although there are challenges in precisely mapping EEG to specific image labels due to the intrinsic noise nature of EEG data and dataset size, these models reliably identify broader categories, proving EEG’s viability for visual concept decoding. Fig 2 presents qualitative examples of predicted and reconstructed images. While there are occasional minor category errors, the model effectively discerns the overarching semantic category and creates corresponding reconstructions, confirming its capability for accurate semantic interpretation from EEG patterns.

4 Discussion

This study aimed to decode and reconstruct visual representations from EEG-recorded brain activity, using deep convolutional neural networks trained on EEG-derived TFD with CLIP-based knowledge distillation. The model demonstrated reliability across most participants but faced challenges distinguishing closely related classes. This approach's potential for non-invasive EEG recordings in brain-computer interfaces is significant, suggesting possible exploitation for artificial vision and innovative neurofeedback experiments (Enriquez-Geppert et al., 2017). However, limitations exist, primarily due to the macroscopic perspective and limited spatial resolution of EEG signals. In the future, integrating EEG with higher-resolution imaging techniques, like fMRI, could enhance image reconstruction detail (Ferrante et al., 2023; Ozcelik & VanRullen, 2023). The model's current limitations include its optimization for a specific set of categories and variability in decoding performance across participants and sessions. Ethical considerations in EEG decoding, particularly around personal perceptual data, are addressed by creating subject-specific models, ensuring consensual and individualized decoding. The study also introduces a training methodology suitable for real-time feedback in models tailored to individuals, with minimal inference time on advanced hardware. Future advancements in deep learning and generative models are expected to further enhance EEG decoding and reconstruction capabilities.

5 Conclusions

In our study, we showcased the capability of deep neural networks combined with generative diffusion models to reconstruct visual experiences from non-invasive EEG recordings. We introduced a novel teacher-student framework where two networks process different yet related data (images and EEG) for the same classification task forcing a representation alignment and we observed that this approach yielded to superior performance in decoding and realistic image reconstruction when combined with a powerful image prior like a pretrained latent diffusion model.

6 Acknowledgements

This work is supported and funded by NEXTGENERATIONEU (NGEU); the Ministry of University and Research (MUR); the National Recovery and Resilience Plan (NRRP); project MNESYS (PE0000006, to NT) - A Multiscale integrated approach to the study of the nervous system in health and disease (DN. 1553 11.10.2022); the MUR-PNRR M4C2I1.3 PE6 project PE00000019 Heal Italia (to NT); the NATIONAL CENTRE FOR HPC, BIG DATA AND QUANTUM COMPUTING, within the spoke "Multiscale Modeling and Engineering Applications" (to NT); the European Innovation Council (Project CROSSBRAIN - Grant Agreement 101070908, Project BRAINSTORM - Grant Agreement 101099355); the Horizon 2020 research and innovation Programme (Project EXPERIENCE - Grant Agreement 101017727). Matteo Ferrante is a Ph.D. student enrolled in the National PhD in Artificial Intelligence, XXXVII cycle, course on Health and Life Sciences, organized by Università Campus Bio-Medico di Roma.

References

Yunpeng Bai, Xintao Wang, Yan-pei Cao, Yixiao Ge, Chun Yuan, and Ying Shan. DreamDiffusion: Generating high-quality images from brain EEG signals. URL `http://arxiv.org/abs/2306.16934`.

Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. Brain decoding: toward real-time reconstruction of visual perception, 2023.

H. M. Bharadwaj, R. B. Wilbur, and J. Siskind. Still an ineffective method with supertrials/erps—comments on "decoding brain representations by multimodal learning of neural activity and visual features". *IEEE Transactions on Pattern Analysis amp; Machine Intelligence*, 45(11): 14052–14054, nov 2023. ISSN 1939-3539. doi: 10.1109/TPAMI.2023.3292062.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Stefanie Enriquez-Geppert, René J. Huster, and Christoph S. Herrmann. Eeg-neurofeedback as a tool to modulate cognition and behavior: A review tutorial. *Frontiers in Human Neuroscience*, 11, 2017. ISSN 1662-5161. doi: 10.3389/fnhum.2017.00051. URL https://www.frontiersin.org/articles/10.3389/fnhum.2017.00051.

Matteo Ferrante, Furkan Ozcelik, Tommaso Boccato, Rufin VanRullen, and Nicola Toschi. Brain captioning: Decoding human brain activity into images and text, 2023.

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 https://www.sciencedirect.com/science/article/pii/S1053811922008758.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

Isaak Kavasidis, Simone Palazzo, Concetto Spampinato, Daniela Giordano, and Mubarak Shah. *Brain2Image*: Converting brain signals into images. In *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1809–1817. ACM. ISBN 978-1-4503-4906-2. doi: 10.1145/3123266.3127907. URL https://dl.acm.org/doi/10.1145/3123266.3127907.

Isaak Kavasidis, Simone Palazzo, Concetto Spampinato, Daniela Giordano, and Mubarak Shah. Brain2image: Converting brain signals into images. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, pp. 1809–1817, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450349062. doi: 10.1145/3123266.3127907. URL https://doi.org/10.1145/3123266.3127907.

Gregory R. Lee, Ralf Gommers, Filip Waselewski, Kai Wohlfahrt, and Aaron O Leary. Pywavelets: A python package for wavelet analysis. *Journal of Open Source Software*, 4(36):1237, 2019. doi: 10.21105/joss.01237. URL https://doi.org/10.21105/joss.01237.

Ren Li, Jared S. Johansen, Hamad Ahmed, Thomas V. Ilyevsky, Ronnie B Wilbur, Hari M Bharadwaj, and Jeffrey Mark Siskind. Training on the test set? an analysis of spampinato et al. [31], 2018.

Ren Li, Jared S. Johansen, Hamad Ahmed, Thomas V. Ilyevsky, Ronnie B. Wilbur, Hari M. Bharadwaj, and Jeffrey Mark Siskind. The perils and pitfalls of block design for eeg classification experiments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):316–333, 2021. doi: 10.1109/TPAMI.2020.2973153.

Furkan Ozcelik and Rufin VanRullen. Brain-diffuser: Natural scene reconstruction from fmri signals using generative latent diffusion, 2023.

Furkan Ozcelik, Bhavin Choksi, Milad Mozafari, Leila Reddy, and Rufin VanRullen. Reconstruction of Perceived Images from fMRI Patterns and Semantic Brain Exploration using Instance-Conditioned GANs, February 2022. URL http://arxiv.org/abs/2202.12692. arXiv:2202.12692 [cs, eess, q-bio].

S. Palazzo, C. Spampinato, I. Kavasidis, D. Giordano, and M. Shah. Generative adversarial networks conditioned by brain signals. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3430–3438. IEEE. ISBN 978-1-5386-1032-9. doi: 10.1109/ICCV.2017.369. URL http://ieeexplore.ieee.org/document/8237631/.

Simone Palazzo, Concetto Spampinato, Joseph Schmidt, Isaak Kavasidis, Daniela Giordano, and Mubarak Shah. Correct block-design experiments mitigate temporal correlation bias in eeg classification, 2020.

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

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

Steffen Schneider, Jin Hwa Lee, and Mackenzie Weygandt Mathis. Learnable latent embeddings for joint behavioural and neural analysis. *Nature*, 617(7960):360–368, May 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06031-6. URL https://doi.org/10.1038/s41586-023-06031-6.

Prajwal Singh, Pankaj Pandey, Krishna Miyapuram, and Shanmuganathan Raman. EEG2image: Image reconstruction from EEG brain signals. URL http://arxiv.org/abs/2302.10121.

C. Spampinato, S. Palazzo, I. Kavasidis, D. Giordano, N. Souly, and M. Shah. Deep learning human mind for automated visual classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4503–4511. IEEE. ISBN 978-1-5386-0457-1. doi: 10.1109/CVPR.2017.479. URL http://ieeexplore.ieee.org/document/8099962/.

Concetto Spampinato, Simone Palazzo, Isaak Kavasidis, Daniela Giordano, Mubarak Shah, and Nasim Souly. Deep learning human mind for automated visual classification, 2019.

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