ENIGMA: A Unified Lightweight EEG-to-Image Model for Multi-Subject Visual Decoding

Reese Kneeland[1] **Wangshu Jiang**[1,2] **Ugo Bruzadin Nunes**[1] **Si Kai Lee**[1]
Paul S. Scotti[3,4] **Arnaud Delorme**[5] **Jonathan Xu**[1]
[1]Alljoined
[2]University of Waterloo
[3]Sophont
[4]Princeton Neuroscience Institute
[5]University of California San Diego

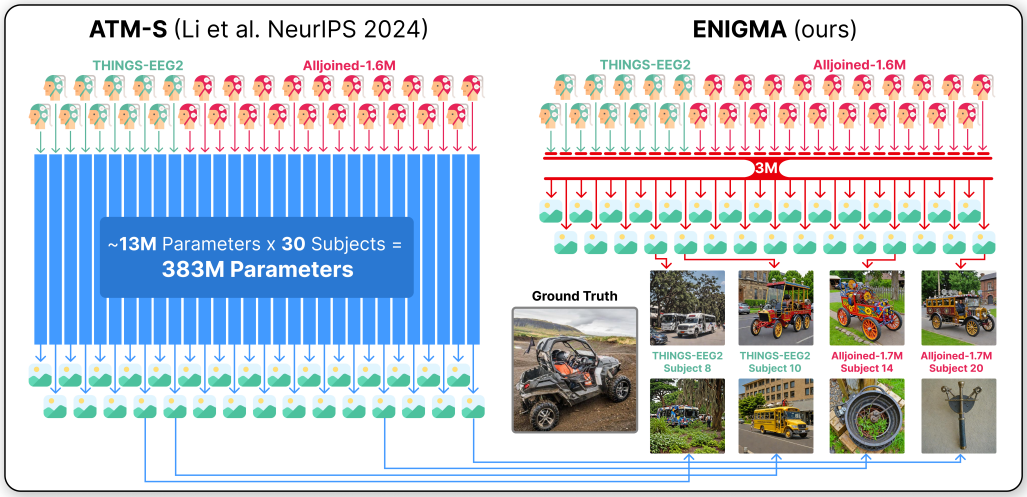

Figure 1: **ENIGMA** (ours) vs ATM-S [1] comparison of model size, methodology, and reconstructions of a seen image from EEG brain activity across subjects from the THINGS-EEG2 [2] (green cap) and Alljoined-1.6M [3] (red cap) datasets.

Abstract

To be practical for real-life applications, models for reconstructing seen images from human brain activity must be effective on affordable scanning hardware, small enough to run locally on accessible computing resources, and easily and consistently deployable across multiple subjects in downstream tasks. To directly address these current limitations, we introduce **ENIGMA**, a multi-subject electroencephalography (EEG)-to-Image decoding model that reconstructs seen images from EEG recordings and achieves state-of-the-art (SOTA) performance on the research-grade THINGS-EEG2 and consumer-grade AllJoined-1.6M benchmarks. **ENIGMA** boasts a simpler architecture and has ∼120x fewer parameters than previous SOTA methods, integrating a set of subject-specific encoder layers with a subject-unified spatio-temporal backbone to map raw EEG signals to a rich visual latent space. We evaluate our approach using a broad suite of image reconstruction metrics that have been standardized in the adjacent field of fMRI-to-Image research, and we describe the first EEG-to-Image study to conduct extensive behavioral evaluations of our reconstructions using human raters. Our simple and robust architecture provides significant performance improvements across both research-grade and consumer-grade EEG hardware, and provides a substantial

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Foundation Models for the Brain and Body.

boost in cross-subject decoding alignment. Finally, we provide extensive ablations to determine the architectural choices most responsible for our performance gains in both single and multi-subject cases across multiple benchmark datasets. Collectively our work provides a substantial step towards the development of practical brain-computer interface applications.

Keywords: EEG; brain decoding; image reconstruction; brain-computer interface; diffusion

1 Introduction

Reconstructing visual experiences from brain activity has long been a goal of both neuroscience and machine learning, and a foundational step for building decoding algorithms for practical brain-computer interface (BCI) applications. While modern generative models and functional magnetic resonance imaging (fMRI) using the Natural Scenes Dataset (NSD) [4, 5] has recently yielded striking reconstructions of seen images using latent diffusion models [6–13], electroencephalography (EEG)-based reconstruction remains challenging due to EEG's low signal-to-noise ratio and spatial resolution. Despite these limitations, EEG remains appealing for real-time BCI applications because of its temporal precision and inexpensive, portable form factor.

Research decoding visual content from EEG recordings spans approaches across classification, retrieval, and image reconstruction. While early studies collected EEG recordings in response to visual stimuli (Spampinato et al. [14]), many contained confounds that made it difficult to decode true semantic content [15], highlighting a need for better data and methods. This critique, along with improvements in EEG preprocessing and experimental design, led to the development of the family of THINGS-EEG datasets as part of the THINGS initiative. Grootswagers et al. [16] released THINGS-EEG (50 subjects, 1,854 concepts) using rapid serial visual presentation, and Gifford et al. [2] further improved upon THINGS-EEG with THINGS-EEG2, emphasizing trial randomization and quality control, which has since become a standard benchmark for EEG vision decoding. On THINGS-EEG2, new methods [17, 1] employing contrastive learning between image and EEG features have achieved significant gains in zero-shot object classification. Alljoined-1.6M [3] extends this paradigm to twice as many subjects and to a consumer-grade EEG hardware setup, providing a new set of tools for developing and evaluating EEG-based visual decoding methods for practical use.

Existing EEG-to-Image decoding research is generally underexplored and spans a wide range of architectural approaches from simple linear mappings Fei et al. [18] to highly complex architectures like ATM-S [1], utilizing combinations of transformer-based brain encoders and multiple diffusion prior training steps. The overlap in these approaches (as well as parallel work in fMRI-to-Image methods) underscore the promise of CLIP-guided diffusion models for reconstructing images from brain activity, however, the drastic discrepancy between the remaining architectural choices highlights the need to examine the effectiveness of these various techniques in the context of broader BCI research and the utility of deploying these models in practical use cases and more affordable form factors. The recent release of the Alljoined-1.6M dataset [3]—designed for evaluating the effectiveness of these methods on affordable hardware—demonstrated that complex architectures such as ATM-S can be fragile in such cases, illustrating the difficulty of translating current research to affordable hardware settings.

Most existing methods for EEG-to-Image decoding [17, 18] require specialized models to be trained for each subject, resulting in a linear increase in parameters as models are trained on additional subjects. While Li et al. [1] does support training a unified model in parallel across multiple subjects, doing so leads to a substantial performance drop, and still requires training separate models for each subject to obtain reasonable performance. Having a subject-unified model that performs consistency and reliably across multiple subjects in parallel at inference time is a very desirable property for decoding models whose outputs are used in downstream BCI applications.

In this paper, we introduce **ENIGMA** (**E**EG **N**eural **I**mage **G**enerator for **M**ulti-subject **A**pplications), a subject-unified model for reconstructing seen images from EEG recordings[1]. Our approach includes several notable contributions to address the gaps in current research identified above:

[1] https://github.com/Alljoined/ENIGMA

(1) ENIGMA produces state-of-the-art performance on both the research-grade THINGS-EEG2 and consumer-grade Alljoined-1.6M EEG benchmark datasets. We are also the first work to provide extensive evaluations of our method using behavioral experiments with human raters. **(2)** Our model is unified across subjects (and datasets), resulting in a ∼120x reduction in the number of parameters (**ENIGMA (Multi-Subject)** vs ATMS (Single-Subject) on THINGS-EEG2 + Alljoined-1.6M) needed to reliably decode images from multiple subjects. **(3)** We conduct detailed analyses of our method and previous approaches on EEG-to-Image decoding: we investigate which aspects of current architectures are most effective across differences in hardware quality/multi-subject use cases, and the alignment consistency of the decoded embedding space in single and multi-subject configurations, for both research-grade and consumer-grade EEG hardware setups.

2 ENIGMA

We designed our model to adhere to several key requirements for practical BCI applications: **(1)** High performance on both research-grade and consumer-accessible EEG hardware. **(2)** A unified model architecture that works across many subjects using only a single model. **(3)** A small, scalable, and efficient design that minimizes model complexity and compute needs.

To meet these requirements, our proposed model, **ENIGMA**, has 4 components (depicted in Figure 2): A set of subject-specific linear encoder layers, a spatio-temporal convolutional neural network to learn the semantic information encoded in the spatial and temporal dimensions of the input signal, an MLP projector to the ViT-H/14 CLIP [19] embedding space, and an image reconstruction module, i.e., SDXL.

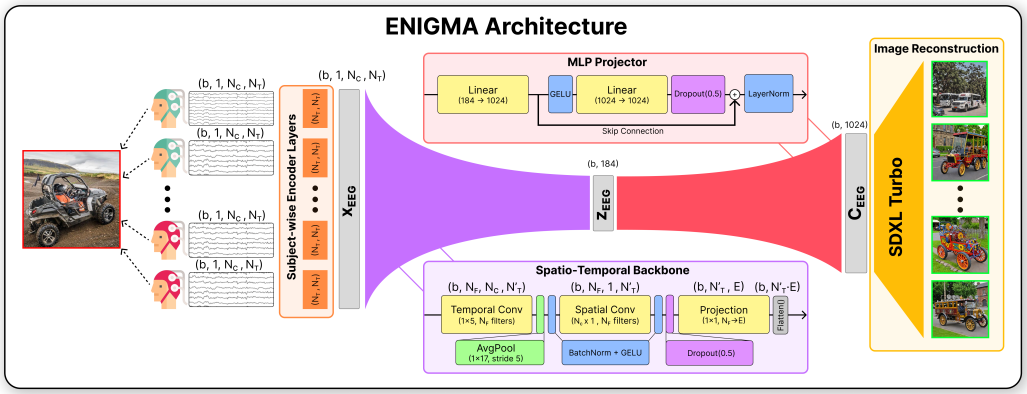

Figure 2: During training, brain activity from each subject is passed through its own subject-specific encoder layer, after which the embeddings x_{EEG} from all subjects are passed through a shared pathway of spatio-temporal convolutions, producing an intermediate latent vector z_{EEG}. This latent is passed through a fully connected MLP projection layer to produce the output c_{EEG} vector. Details of these procedures are provided in Section 2.2.

2.1 Datasets

THINGS-EEG2 [2] is the current public benchmark for EEG-based visual decoding. It contains 64-channel ActiChamp (costing ∼$60,000) recordings recorded at 250Hz from 10 participants viewing 16740 unique images in a rapid serial visual-presentation paradigm. 200 of these images are designated as the testing set and are repeated 80 times each, while the remaining 16540 images in the training set are repeated 4 times each, for a total of ∼820k trials across the whole dataset

Alljoined-1.6M [3] is a follow-up corpus of EEG responses to visual stimuli collected with a much cheaper 32-channel Emotiv Flex 2 gel headset (∼$2,200) at 250Hz comprising the same stimuli and experimental paradigm as THINGS-EEG2, across 20 new subjects, for a total of ∼1.6M trials.

Our preprocessing steps are in A.1. When reproducing other methods, we follow their preparation and preprocessing steps.

2.2 Architecture

Subject-wise Encoder Layers To account for systematic differences between subjects, we learn a set of subject-specific fully-connected linear encoder layers $W_s \in \mathbb{R}^{N_T \times N_T}$ that output a common aligned representation x_{EEG} across subjects. This alignment step ensures that downstream modules learn a shared CLIP embedding across subjects, which reduces the number of parameters needed by our multi-subject model.

Spatio-Temporal Backbone Next, the aligned EEG data x_{EEG} is passed through an embedding module that applies convolutions over the temporal and spatial dimensions of our data.

MLP Projector After obtaining the z_{EEG} feature vector, we use a projection head to predict a c_{EEG} embedding mapped to the final CLIP ViT-H/14 latent space $D = 1024$.

Image Reconstruction To generate images from the EEG embedding c_{EEG}, we leverage Stable Diffusion XL Turbo (SDXL) [20], and its associated CLIP ViT-H/14 image-prompt adapter (IP-Adapter) [21].

Loss Function Following Li et al. [1], we align the EEG embedding c_{EEG} to the CLIP ViT-H/14 image embedding of the stimulus image $f_{\text{CLIP}}(\text{image}) \in \mathbb{R}^{1024}$ by minimizing the Mean-Squared Error (MSE) between the two and regularizing with the InfoNCE contrastive loss [22, 19]. The former matches the EEG embedding to its corresponding image embedding in CLIP latent space, and the latter ensures that the embedding retains relevant directional semantics within the CLIP manifold, while learning to discard the subject and session-specific information. The relative weight of these two losses is modulated by $\lambda = 0.5$, making our overall loss function:

$$\mathcal{L} = \text{MSE}(c_{\text{EEG}}, f_{\text{CLIP}}(\text{image})) + \lambda\, \text{InfoNCE}(c_{\text{EEG}}, \text{norm}(f_{\text{CLIP}}(\text{image}))$$

3 Results

We evaluate **ENIGMA** against the only available EEG-to-Image baselines, Perceptogram [18] and ATM-S [1], and report results on two of the most prominent benchmarks: THINGS-EEG2 [2] and Alljoined-1.6M [3].

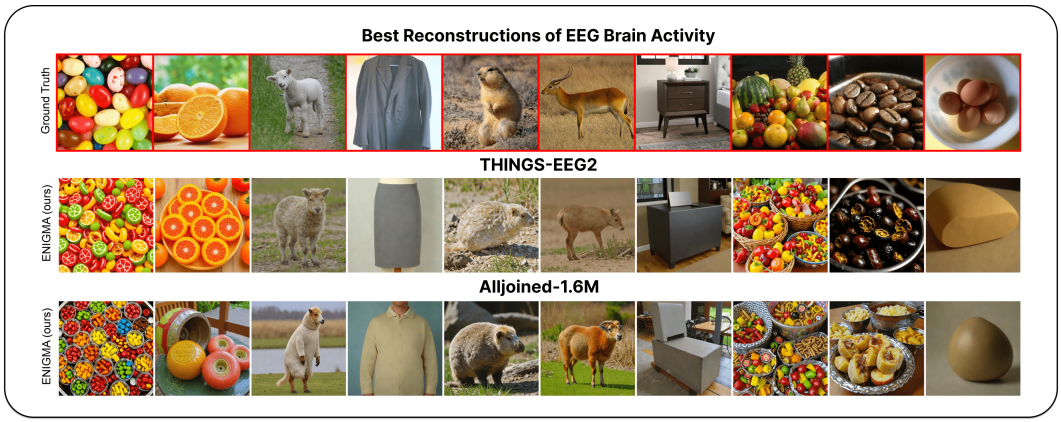

Figure 3: Qualitative comparison of reconstruction methods on seen stimuli from THINGS-EEG2 and Alljoined-1.6M.

Figure 3 presents a set of the best reconstructed images from EEG. A qualitative comparison to previous work can be found in Appendix A.7. These examples illustrate typical outcomes: our reconstructions generally capture the correct high-level object, e.g., oranges, sheep, furniture, etc, and reproduce a lot of the semantic content contained in the original images.

Table 1 summarizes the quantitative performance of **ENIGMA** and baseline methods on both benchmarks in single and multi-subject configurations. For all methods, we output 10 reconstructions per test sample from each method and report averaged image feature metrics across them. For multi-subject configurations (our primary evaluation target) **ENIGMA** achieves the best scores on all metrics, indicating our subject-specific layers are enabling our model to scale across subjects from both datasets. For single-subjects, our model still provides state-of-the-art (SOTA) performance on

Method	Model Properties	Low-Level		High-Level						Retrieval			Human Raters
	# of Parameters ↓	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓	Top-1 ↑	Top-5 ↑	Top-10 ↑	Ident. Acc. ↑
THINGS-EEG2													
ENIGMA (Multi-Subject)	**3,175,392**	**0.162**	**0.413**	**80.49%**	**86.54%**	**73.45%**	**77.34%**	**0.878**	**0.553**	**22.55%**	**50.75%**	**64.05%**	**82.03%**
ATM-S (Multi-Subject)	12,815,311	0.072	0.403	57.09%	58.99%	52.86%	55.04%	0.963	0.663	16.20%	45.10%	62.20%	56.82%
ENIGMA (Single-Subject)	**14,052,420**	0.159	0.422	81.89%	**88.34%**	**75.09%**	**78.90%**	**0.870**	**0.546**	27.60%	59.35%	71.15%	**83.06%**
ATM-S (Single-Subject)	128,153,110	0.136	0.392	73.85%	80.83%	67.56%	71.28%	0.909	0.601	**30.15%**	**60.15%**	**73.60%**	77.14%
Perceptogram (Single-Subject)	4,731,924,800	**0.247**	**0.431**	**85.46%**	88.03%	70.40%	71.98%	0.902	0.581	–	–	–	79.17%
Alljoined-1.6M													
ENIGMA (Multi-Subject)	**3,175,392**	**0.086**	0.375	**63.97%**	**70.30%**	**61.49%**	**64.59%**	**0.929**	**0.612**	**6.00%**	**18.85%**	**28.80%**	**68.07%**
ATM-S (Multi-Subject)	12,765,711	0.068	**0.427**	53.49%	53.36%	50.72%	51.46%	0.965	0.668	0.73%	4.13%	7.55%	52.18%
ENIGMA (Single-Subject)	**27,112,840**	0.079	**0.416**	63.62%	67.84%	**59.57%**	**62.91%**	**0.942**	**0.620**	**6.00%**	**16.25%**	**25.35%**	**65.43%**
ATM-S (Single-Subject)	255,314,220	0.090	0.374	55.91%	58.25%	54.07%	56.25%	0.960	0.673	0.50%	2.00%	5.00%	60.31%
Perceptogram (Single-Subject)	9,463,849,600	**0.094**	0.401	**67.36%**	**69.28%**	58.18%	59.94%	0.945	0.637	–	–	–	62.00%

Table 1: Comparison of EEG-to-Image reconstruction models on the THINGS-EEG2 and Alljoined-1.6M datasets via image similarity metrics. Parameter counts are computed by adding up the number of parameters used to decode all subjects in each dataset (10 subjects for THINGS-EEG2, 20 for Alljoined-1.6M). Details on the human identification accuracy metric are provided in Section 3.1. For the number of parameters, EffNet-B, and SwAV, lower is better. For all other metrics, higher is better. Bold indicates best performance, and underlines second-best performance. Additional details on the metrics are in Appendix A.5.

the majority of metrics. We note that although many of the listed metrics are often used as a proxy for human judgment, research has established that these metrics do not closely approximate or align with human assessments of content [23] or quality [24]. Thus, we also provide human identification accuracy scores in the "Human Raters" section, discussed further in Section 3.1.

3.1 Human Behavioral Evaluations

For brain decoding models to be used in BCI applications, users, scientists, and clinicians need to be able to meaningfully interpret the outputs of such models. Thus, human judgments of the quality of EEG-to-Image reconstructions is an important performance metric. In light of this, we conducted a large-scale online behavioral experiment where human raters (n = 545) assessed the quality of the reconstructions. Detailed experiment protocols are in Appendix A.10.

We asked human raters to perform a 2-alternative forced choice judgment about whether a reconstruction was more similar to the ground truth image than a randomly selected reconstruction of a different stimulus sampled from the same reconstruction method, dataset, and subject. Accurately matching a reconstruction to its corresponding stimulus image is a minimum requirement for confirming that the reconstruction contains meaningful content. Our results (Table 1) confirm **ENIGMA** as SOTA for the human identification accuracy of EEG-to-Image reconstructions in all cases.

4 Discussion

Here we introduce **ENIGMA**, a subject-unified EEG-to-Image reconstruction model, which achieves SOTA results on multiple datasets while drastically reducing the number of model parameters relative to existing approaches, and enabling scalable multi-subject deployment. By combining subject-specific encoder layers, an efficient spatio-temporal encoding module, and a fast and lightweight image generator, **ENIGMA** is able to reliably and consistently reconstruct visual stimuli from multi-subject EEG recordings with unprecedented semantic accuracy. Our analysis of existing techniques in Appendix A.3 highlights the balance between optimizing for performance with complex, finely tuned architectures, and optimizing for robustness and flexibility to consumer-grade EEG data recorded across multiple participants. **ENIGMA** aims to strike a balance between these desiderata, and lays a framework for future research in this space.

Despite fMRI's dominance in prior neural image reconstruction work, limited accessibility, high relative cost, and poor temporal resolution make fMRI impractical for real-world BCI applications. **ENIGMA** marks a significant step toward deployable EEG decoding systems—potentially enabling applications like at-home assistive communication tools or rapid stimulus decoding in clinical and research settings. The success of **ENIGMA** demonstrates that the semantic bottleneck traditionally attributed to EEG recordings is not entirely an inherent data quality limitation, but a function of suboptimal encoding and decoding strategies. Compared to fMRI, EEG provides vastly inferior

spatial resolution, yet it remains a much more practical modality for downstream BCI applications, as fMRI is not possible to deploy outside of a lab. Our results suggest that with such lightweight, subject-adaptive architectures in place, EEG has plenty of practical utility as an interface for applications of brain decoding technology.

4.1 Ethical Considerations

Research aimed at decoding cognitive states is rapidly growing in scope and capability. While these endeavors promise clear downstream benefits, they also raise serious questions about broader societal implications and their potential for misuse. Hence, the importance of developing an ethical framework for the application of brain decoding devices that rigorously safeguards users' data and ensures that the technology is deployed transparently, responsibly, and for the benefit of humankind [25].

References

[1] Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, and Quanying Liu. Visual Decoding and Reconstruction via EEG Embeddings with Guided Diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[2] Alessandro T. Gifford, Kshitij Dwivedi, Gemma Roig, and Radoslaw M. Cichy. A large and rich EEG dataset for modeling human visual object recognition. *NeuroImage*, 264:119754, 2022. doi: 10.1016/j.neuroimage.2022.119754.

[3] Anonymous. Alljoined-1.6m: A million-trial eeg-image dataset for evaluating affordable brain-computer interfaces. In Review, 2025.

[4] Emily J. Allen, Ghislain St-Yves, Yihan Wu, Jesse L. Breedlove, Jacob S. Prince, Logan T. Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, J. Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, January 2022. ISSN 1097-6256, 1546-1726. doi: 10.1038/s41593-021-00962-x. URL https://www.nature.com/articles/s41593-021-00962-x.

[5] Reese Kneeland, Paul S. Scotti, Ghislain St-Yves, Jesse Breedlove, Kendrick Kay, and Thomas Naselaris. NSD-Imagery: A benchmark dataset for extending fMRI vision decoding methods to mental imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025.

[6] Furkan Ozcelik and Rufin VanRullen. Natural scene reconstruction from fMRI signals using generative latent diffusion. *Scientific Reports*, 13, 2023. URL https://api.semanticscholar.org/CorpusID:260439960.

[7] Paul Steven Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Cohen Ethan, Aidan James Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth Norman, and Tanishq Mathew Abraham. Reconstructing the mind's eye: fMRI-to-image with contrastive learning and diffusion priors. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=rwrblCYb2A.

[8] Paul S. Scotti, Mihir Tripathy, Cesare Kadir Torrico Villanueva, 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 *Proceedings of the 41st International Conference on Machine Learning*, 2024.

[9] 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*, pages 14453–14463, 2023.

[10] Yu Takagi and Shinji Nishimoto. Improving visual image reconstruction from human brain activity using latent diffusion models via multiple decoded inputs, 2023.

[11] Reese Kneeland, Jordyn Ojeda, Ghislain St-Yves, and Thomas Naselaris. Brain-optimized inference improves reconstructions of fMRI brain activity, December 2023. URL http://arxiv.org/abs/2312.07705. arXiv:2312.07705 [cs, q-bio].

[12] Reese Kneeland, Jordyn Ojeda, Ghislain St-Yves, and Thomas Naselaris. Reconstructing seen images from human brain activity via guided stochastic search. In *Conference on Cognitive Computational Neuroscience*, 2023. doi: 10.32470/CCN.2023.1672-0. URL https://2023.ccneuro.org/view_paper1337.html?PaperNum=1672.

[13] Reese Kneeland, Jordyn Ojeda, Ghislain St-Yves, and Thomas Naselaris. Second Sight: Using brain-optimized encoding models to align image distributions with human brain activity, June 2023. URL http://arxiv.org/abs/2306.00927. arXiv:2306.00927 [cs, q-bio].

[14] Carlo Spampinato, Sebastiano Palazzo, Ignazio Kavasidis, Daniele Giordano, Nada Souly, and Mubarak Shah. Deep Learning Human Mind for Automated Visual Classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6809–6818, 2017.

[15] Ren Li, Jared S. Johansen, Hamad Ahmed, Thomas V. Ilyevsky, Ronnie B. Wilbur, Hari M. Bharadwaj, and Jeffrey M. Siskind. Training on the test set? An analysis of Spampinato et al.'s EEG image classification method. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2020. *(Early Access) arXiv:1812.07697*.

[16] 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(3), 2022. doi: 10.1038/s41597-021-01102-7.

[17] Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding Natural Images from EEG for Object Recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

[18] Teng Fei, Abhinav Uppal, Ian Jackson, Srinivas Ravishankar, David Wang, and Virginia R. de Sa. Perceptogram: Reconstructing Visual Percepts from EEG. *arXiv preprint arXiv:2404.01250*, 2024. (extended version with additional analyses).

[19] 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, February 2021. URL http://arxiv.org/abs/2103.00020. arXiv:2103.00020 [cs].

[20] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pages 87–103. Springer, 2024.

[21] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. *arXiv preprint arXiv:2308.06721*, 2023.

[22] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. URL https://arxiv.org/abs/1807.03748.

[23] Pawan Sinha and Richard Russell. A perceptually based comparison of image similarity metrics. *Perception*, 40(11):1269–1281, 2011. doi: 10.1068/p7063. URL https://doi.org/10.1068/p7063. PMID: 22416586.

[24] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=G5RwHpBUv0.

[25] Emma C. Gordon and Anil K. Seth. Ethical considerations for the use of brain–computer interfaces for cognitive enhancement. *PLOS Biology*, 22(10):1–15, 10 2024. doi: 10.1371/journal.pbio.3002899. URL https://doi.org/10.1371/journal.pbio.3002899.

[26] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A. Engemann, Daniel Strohmeier, Christian Brodbeck, and et al. Meg and eeg data analysis with mne-python. *Frontiers in Neuroscience*, 7:267, 2013. doi: 10.3389/fnins.2013.00267.

[27] Nicholas A Badcock, Petroula Mousikou, Yatin Mahajan, Peter De Lissa, Johnson Thie, and Genevieve McArthur. Validation of the emotiv epoc® eeg gaming system for measuring research quality auditory erps. *PeerJ*, 1:e38, 2013.

[28] Nikolas S Williams, Genevieve M McArthur, Bianca de Wit, George Ibrahim, and Nicholas A Badcock. A validation of emotiv epoc flex saline for eeg and erp research. *PeerJ*, 8:e9713, 2020.

[29] Matthias Guggenmos, Philipp Sterzer, and Radoslaw M Cichy. Multivariate pattern analysis for meg: A comparison of dissimilarity measures. *NeuroImage*, 173:434–447, 2018.

[30] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[31] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, April 2004. ISSN 1941-0042. doi: 10.1109/TIP.2003.819861. Conference Name: IEEE Transactions on Image Processing.

[32] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[33] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015. URL http://arxiv.org/abs/1512.00567.

[34] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 2019. URL http://proceedings.mlr.press/v97/tan19a.html.

[35] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *CoRR*, abs/2006.09882, 2020. URL https://arxiv.org/abs/2006.09882.

A Appendix

A.1 Data Processing and Format

Raw EEG was stored in standard `.edf` files and pre-processed with MNE-Python [26]. The Emotiv firmware applies a dual 50/60 Hz notch by default, effectively attenuating frequencies above 43 Hz, so we added a 0.5 Hz high-pass and an extra 60 Hz notch to suppress residual line noise. We epoched continuous recordings from -200 ms to 1000 ms relative to image onset. Synchronisation glitches in the Emotiv trigger stream led us to discard 0.55–1.12% of trials, which was comparable to exclusion rates reported in earlier Emotiv evaluations [27, 28]. Epochs were then baseline-corrected to the pre-stimulus window and resampled to 250 Hz to match the ATM-S benchmark [1]. Finally, we performed multivariate noise normalization [29] where we whiten input data to improve signal-to-noise ratio. Note that we estimate the whitening matrix only on the training partition to avoid data contamination. This yielded samples $x_{\text{EEG}} \in \mathbb{R}^{N_C \times N_T}$ with $N_C = 64$ channels for THINGS-EEG2, $N_C = 32$ channels for Alljoined-1.6M, and $N_T = 250$ time points for both datasets. For multi-subject models, all models require the same channel count across all subjects, and so for these models we subsample THINGS-EEG2 to the same 32 channels present in Alljoined-1.6M. For an analysis of this step on performance, see Appendix A.8.

A.2 Architecture Details

Spatio-Temporal Backbone Our convolution module treats multichannel EEG as an "image" of shape $1 \times N_C \times N_T$. A temporal 2D convolution (kernel $(1, 5)$, $N_F = 40$ feature maps) is followed by average pooling over time (kernel $(1, 17)$, stride 5), batch normalization (BN), and a GELU non-linearity:

$$h_1 = \text{GELU}\Big(\text{BN}\big(\text{AvgPool}_{1\times 17,5}\big(\text{Conv2D}^{N_F}_{1\times 5}(x_{EEG})\big)\big)\Big), \qquad h_1 \in \mathbb{R}^{N_F \times N_C \times N'_T},$$

where $N'_T = \left\lfloor \frac{N_T - 5 + 1 - 17}{5} \right\rfloor + 1$ (i.e., $N_T = 250 \Rightarrow N'_T = 46$). Next, a spatial convolution (kernel $(N_C, 1)$, N_F) integrates information across channels, followed by BN and GELU yields:

$$h_2 = \text{GELU}\Big(\text{BN}\big(\text{Conv2D}^{N_F}_{N_C\times 1}(h_1)\big)\Big), \qquad h_2 \in \mathbb{R}^{N_F \times 1 \times N'_T}.$$

We apply dropout [30] ($p = 0.5$) for regularization and project the features to an embedding dimension $E_p = 4$ with a 1×1 convolution. Finally, we flatten the output $\in \mathbb{R}^{E \times 1 \times N'_T}$ to a sequence $z_{EEG} \in \mathbb{R}^{N'_T \cdot E}$ to obtain 184 features per trial.

MLP Projector The prediction head of our model is a feed-forward MLP network with a skip connection: a linear layer from 184 to 1024, followed by GELU and dropout, then another linear layer, and finally layer normalization. The residual is added to the output of the second linear layer before normalization, to help stabilize training and allow the model to refine the initial linear projection with non-linear adjustments. The module outputs the EEG embedding $c_{\text{EEG}} \in \mathbb{R}^{1024}$.

Image Reconstruction Our model utilizes SDXL Turbo with an IP-Adapter. The IP-Adapter is a lightweight module inserted into SDXL's cross-attention layers, which enables an image embedding to steer image generation alongside an optional text prompt. We optimize our model to predict the CLIP ViT-H/14 image embeddings expected by SDXL Turbo's IP-adapter as input. Formally, SDXL solves:

$$x_T \sim \mathcal{N}(0, I),$$
$$x_{t-1} = f_\theta(x_t, c_{\text{text}}, c_{\text{EEG}}, t) + \text{noise},$$

for $t = T, T-1, \ldots, 0$, where c_{text} is the text context (which in our case is an unconditional placeholder embedding) and c_{EEG} is our injected EEG image embedding. We run the diffusion for 4 inference steps, which is standard for this version of SDXL.

Loss Function Unlike Li et al. [1], we chose not to normalize $f_{\text{CLIP}}(\text{image})$ in the MSE component to ensure that the learned c_{EEG} respects the geometry of the CLIP embedding space. Doing so negates the need for the secondary diffusion prior training stage in Li et al. [1] that was used to learn the magnitude of the CLIP embedding. This intuition is confirmed by our ablation analyses in Section A.3, which shows that the inclusion of the diffusion prior negatively impacts reconstruction performance for our architecture.

Training Details Training is performed in FP32 on an RTX 3090 GPU with 24GB of VRAM, using AdamW optimizer with learning rate $3e-4$. We train for 150 epochs on the training split of both THINGS-EEG2 and Alljoined-1.6M simultaneously (30 subjects, ∼2M training trials). The model takes 21 hours to train across all 30 subjects, and 45m to train for a single subject on an RTX 3090 GPU. We note that in practice our model could also be trained on GPUs with as little as 8GB of VRAM.

A.3 Ablation/Scaling Analyses

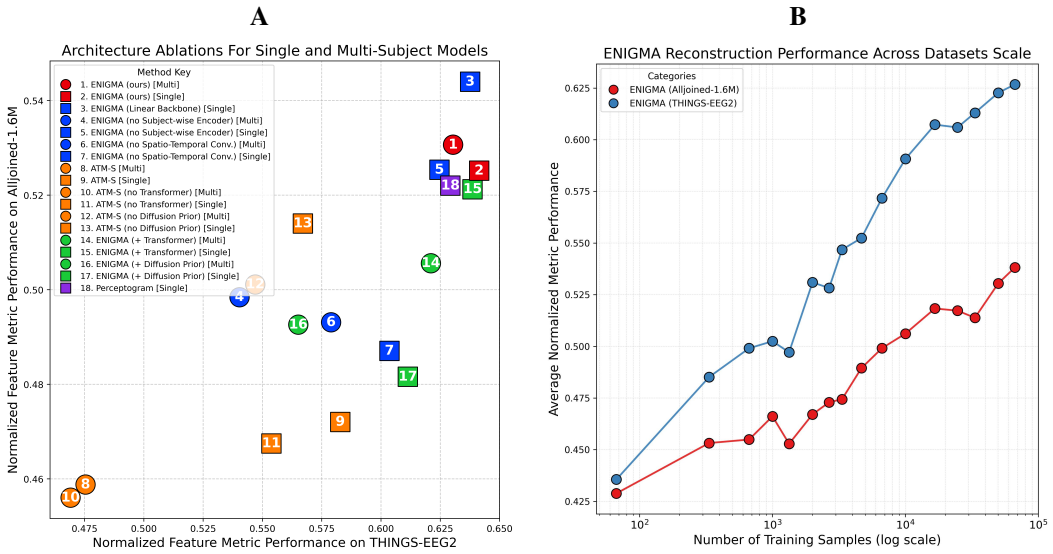

Figure 4: **(A)** Ablation analyses: model variants (numbered icons) in single (square) and multi-subject (circle) configurations under each ablation type (color) are assessed via the normalized average of all feature metrics (Table 1), with THINGS-EEG2 performance on the x-axis and Alljoined-1.6M performance on the y-axis. **(B)** Scaling analysis of **ENIGMA** performance on the THINGS-EEG2 and Alljoined-1.6M datasets. The number of training samples are plotted on a log-scale X-axis, and the normalized average of feature metrics presented in Table 1 is plotted on the Y-axis.

To rigorously evaluate model architectures suited for decoding mental images and critically examine why our method succeeds in multi-subject contexts where ATM-S breaks down, we conducted extensive ablations on key design choices in our method. Colored numeric identifiers refer to the ablation results in Figure 4A. **ENIGMA** is set as (1), [2].

ENIGMA Modules. We find using a linear backbone [3] remains a highly effective way to decode semantic information from brain activity, although we note that linear models do not provide any of the parameter efficiency or multisubject benefits of our ENIGMA architecture. We also find that eliminating the subject-wise encoder (4), [5] harms performance disproportionately in multi-subject contexts, showing that the module specifically drives cross-participant generalization. Removing the spatio-temporal convolution stack decreases accuracy in all contexts (6), [7], confirming that joint space-time feature extraction is essential for capturing the semantic information encoded in brain activity.

ATM-S Modules. While ATM-S's (8), [9] diffusion prior slightly improves its performance on THINGS-EEG2 [13], we find in our ablation analysis that it significantly harms performance on the cheaper EEG hardware in Alljoined-1.6M, and on both datasets in multi-subject contexts (12). We observe a similar outcome with the transformer-based encoder in ATM-S (10), [11], where it only significantly improves performance on THINGS-EEG2 in single-subject contexts. We interpret these results to support our hypothesis that these complex architectural elements are poorly suited to decoding tasks on the lower-SNR data in Alljoined-1.6M, and that these architectures are not well suited for capturing nuanced differences between multiple subjects in a unified architecture.

ENIGMA + ATM-S Modules. To evaluate whether the architectural modules introduced by Li et al. [1] would be beneficial to **ENIGMA**, we grafted ATM-S's transformer-based encoder and diffusion prior training stage onto **ENIGMA** (14),[15],(16),[17]. We find that both of these modules harm performance on both datasets in both single and multi-subject contexts, highlighting how our simple

and robust design captures all of the necessary information encoded in brain activity without needing additional expensive modules.

A.4 Scaling Analysis

As shown in Figure 4B, the reconstruction performance of **ENIGMA** increases log-linearly with the number of training samples, with no evident saturation on either dataset. Performance improves more quickly on THINGS-EEG2 that was collected on much more expensive hardware. Such divergence suggests that while sheer data volume does reliably boosts accuracy, the quality of recording hardware significantly accelerates learning efficiency and leaves headroom for further gains. As highlighted with the release of Alljoined-1.6M [3], this difference in scaling efficiency between EEG hardware quality is a key limitation to overcome for practical BCI applications.

A.5 Additional Details on Evaluation Metrics

We use the following image similarity metrics:

- PixCorr is the pixel-level correlation between the ground-truth images and reconstructed images.
- SSIM is the structural similarity index metric [31].
- AlexNet(2) and AlexNet(5) are the 2-way comparisons (2WC) of layers 2 and 5 of AlexNet [32].
- CLIP is the 2WC of the output layer of the CLIP ViT-L/14 Vision model [19].
- Incep is the 2WC of the last pooling layer of InceptionV3 [33].
- Eff and SwAV are distance metrics gathered from EfficientNet-B13 [34] and SwAV-ResNet50 [35] models.

For the metrics in Table 1, a two-way comparison (2WC) evaluates whether the feature embedding of the stimulus image is more similar to the feature embedding of the target reconstruction, or the feature embedding of a randomly selected "distractor" reconstruction, where the score is the percent of correctly identified target reconstructions across a pool of "distractors". Our 2WC metrics, calculated using reconstructions of the 199 other test-set stimuli as "distractors", have a notably different chance threshold from 2WC metrics presented in reconstruction papers that perform evaluations using a test set with a different number of "distractors", such as the shared1000 test set of NSD [4], and are thus not directly comparable. All metrics in Table 1 were calculated and averaged across 10 images sampled from the output distribution of each method using a random seed. All metrics in Table were calculated on our reproduction of other methods using their open source code, and might differ slightly from metrics reported in the original papers due to our implementation of the metrics we calculated.

A.6 Statistical Significance of Evaluation Metrics

Method	Low-Level				High-Level				Human Raters
	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓	Ident. Acc. ↑
THINGS-EEG2									
ENIGMA (Multi-Subject)	±0.0014	±0.0014	±0.15%	±0.12%	±0.20%	±0.19%	±0.0008	±0.0008	±0.89%
ATM-S (Multi-Subject)	±0.0009	±0.0010	±0.20%	±0.20%	±0.21%	±0.21%	±0.0004	±0.0006	±1.15%
ENIGMA (Single-Subject)	±0.0014	±0.0014	±0.14%	±0.11%	±0.20%	±0.18%	±0.0008	±0.0008	±0.87%
ATM-S (Single-Subject)	±0.0013	±0.0013	±0.17%	±0.15%	±0.21%	±0.20%	±0.0007	±0.0008	±0.97%
Perceptogram (Single-Subject)	±0.0014	±0.0015	±0.12%	±0.11%	±0.20%	±0.20%	±0.0007	±0.0007	±0.94%
Alljoined-1.6M									
ENIGMA (Multi-Subject)	±0.0007	±0.0008	±0.14%	±0.13%	±0.15%	±0.15%	±0.0005	±0.0005	±0.77%
ATM-S (Multi-Subject)	±0.0007	±0.0007	±0.14%	±0.15%	±0.15%	±0.15%	±0.0003	±0.0004	±0.82%
ENIGMA (Single-Subject)	±0.0007	±0.0009	±0.14%	±0.14%	±0.15%	±0.15%	±0.0004	±0.0005	±0.78%
ATM-S (Single-Subject)	±0.0007	±0.0008	±0.14%	±0.15%	±0.15%	±0.15%	±0.0004	±0.0005	±0.80%
Perceptogram (Single-Subject)	±0.0008	±0.0010	±0.13%	±0.13%	±0.15%	±0.15%	±0.0004	±0.0005	±0.79%

Table 2: Standard error measurements for evaluation metrics of EEG-to-Image reconstruction models evaluated on the THINGS-EEG2 and Alljoined-1.6M datasets. Values correspond to the standard error spread of values in Table 1 in the manuscript.

A.7 Reconstruction Comparison

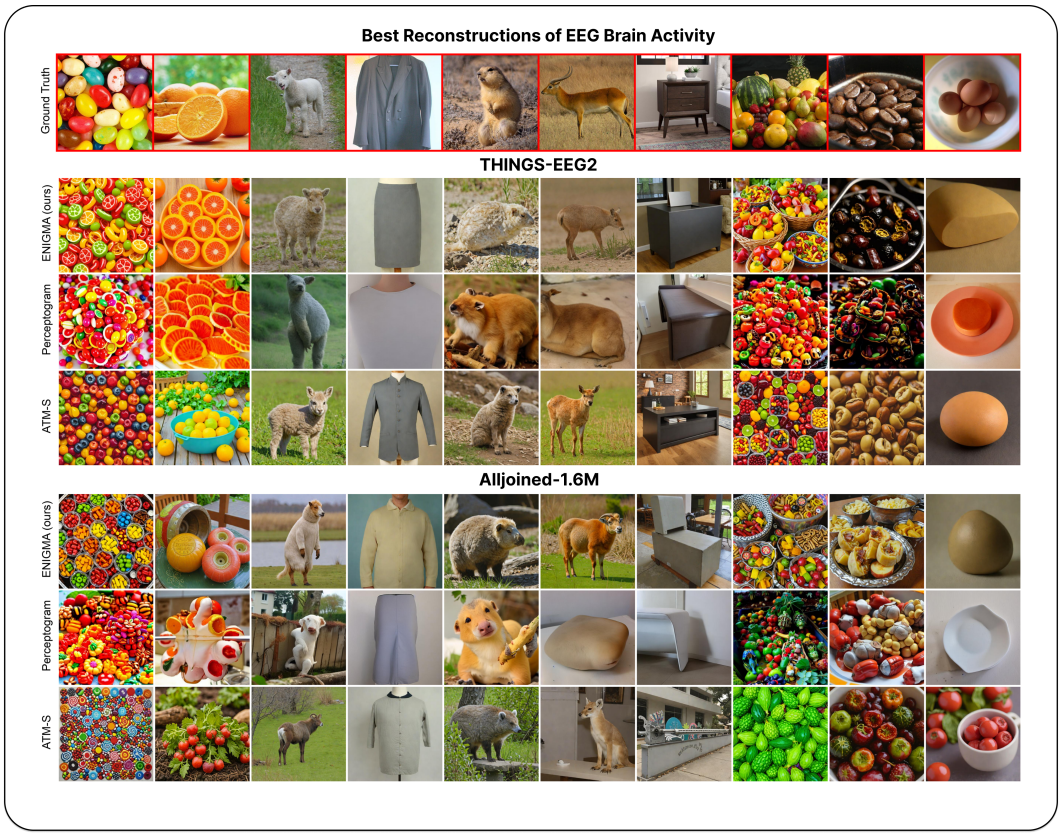

Figure 5: Qualitative comparison of reconstruction methods on seen stimuli from THINGS-EEG2 and Alljoined-1.6M. Reconstructions selected are the outputs sampled from each method and stimulus with the highest scores on all of the image feature metrics in Table 1.

A.8 Channel Count Ablation Analysis

A commonly-asked question is the number of channels needed to obtain high quality reconstructions from EEG-to-Image reconstruction models. We analyzed how the number of channels affects

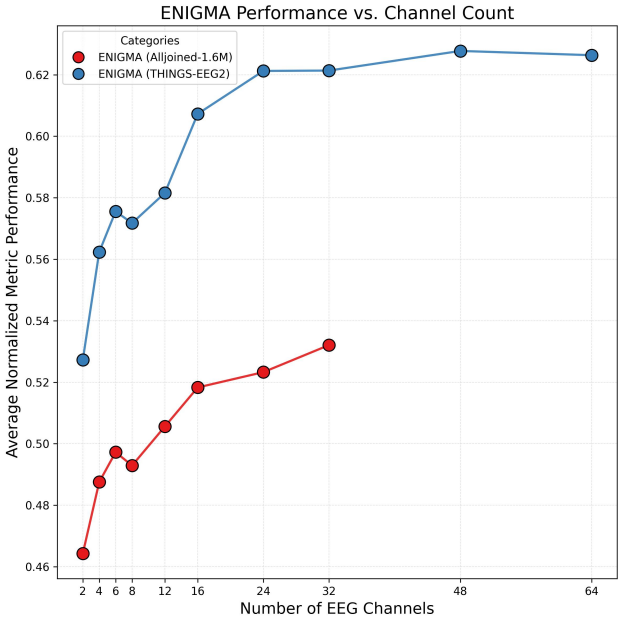

Figure 6: Channel count analysis of model performance for each dataset. The number of channels in each dataset was progressively reduced, while the remaining channels focus primarily on occipital cortex. The Y axis is plotted the same as Fig. 4B.

decoding performance using **ENIGMA**, and explored whether this contributed significantly to differences in performance between performance on the two benchmark datasets. We sub-sampled varying numbers of channels from both datasets, while retaining a focus on covering occipital cortex. We find that while performance did drop with fewer channels, channel count is not the most significant factor accounting for the performance difference between the datasets, and performance starts to drop off after 24 channels for both datasets. This suggest that it might be possible to achieve reasonable decoding performance with fewer than 32 channels.

A.9 Subject-wise Correlation Plots

A desirable quality of brain decoding models is to having a model that produces consistent outputs across multiple subjects viewing the same image. To explore this, we plotted the mean correlation of the predicted c_{EEG} vectors across subjects from both datasets in Figure 7. The bottom row of the figure plots the mean correlation for single-subject models. We note that when training individual models for each subject, models generally have poor subject-wise correlations, i.e., many off-diagonal entries have correlation ≤ 0.5, demonstrating that single-subject models are learning embeddings that are mostly subject-independent. While both the **ENIGMA** and ATM-S models appear highly correlated in the multi-subject case, the ATM-S model is actually almost entirely "collapsed", i.e., every test image produces a near identical reconstruction, resulting in near-chance image identification scores. The top right subplot visualizes both subject-to-subject correlations and reconstruction performance on the X and Y axis respectively, with our multi-subject **ENIGMA** architecture landing in the desirable top right corner.

In Figure 8, we plot the mean correlation of the x_{EEG} embeddings produced by the subject-specific modules of **ENIGMA** and ATM-S with the raw EEG recordings as baseline. Notice the lack of structure in the x_{EEG} embeddings for the single-subject configurations of **ENIGMA** and ATM-S as compared to the raw EEG recordings (which already surfaces inherent differences between the subjects from THINGS-EEG2 and Alljoined-1.6M). This is expected as each single-subject model is subject-specific, hence the embeddings should not be correlated. We observe that x_{EEG} embeddings for multi-subject configurations of **ENIGMA** and ATM-S have much more obvious correlation structures both for within-dataset subjects and across-dataset subjects. In particular, the linear transform of multi-subject **ENIGMA** accentuates existing correlation structure in the raw EEG recordings for within-dataset and across-dataset subjects, i.e., increases the magnitude of positive

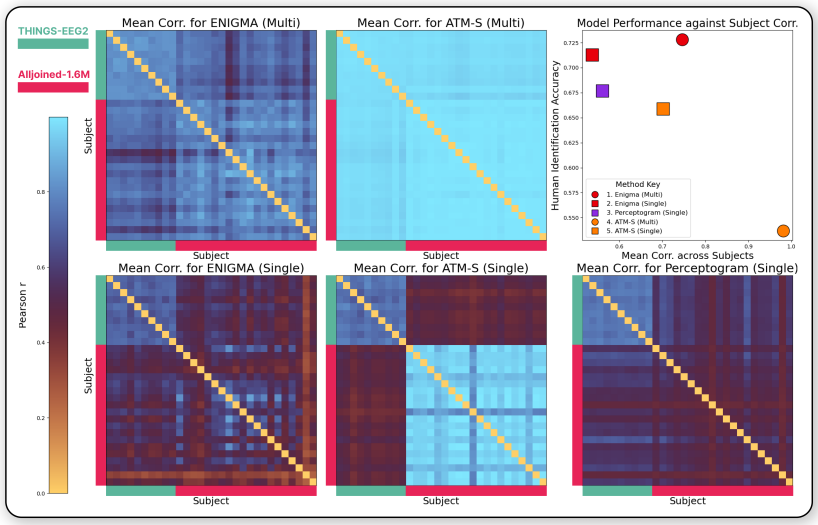

Figure 7: Subject-to-subject Pearson correlation matrices of predicted CLIP embeddings from different subjects in THINGS-EEG2 (green) and Alljoined-1.6M (red). Plots are displayed for all methods in Table 1 in both single and multi-subject configurations. The top right figure plots the human identification accuracy from the behavioral experiment in Section 3.1 (averaged across THINGS-EEG2 and Alljoined-1.6M subjects) against the average of the off-diagonals of the correlation matrices. Chance performance for human identification is 0.5.

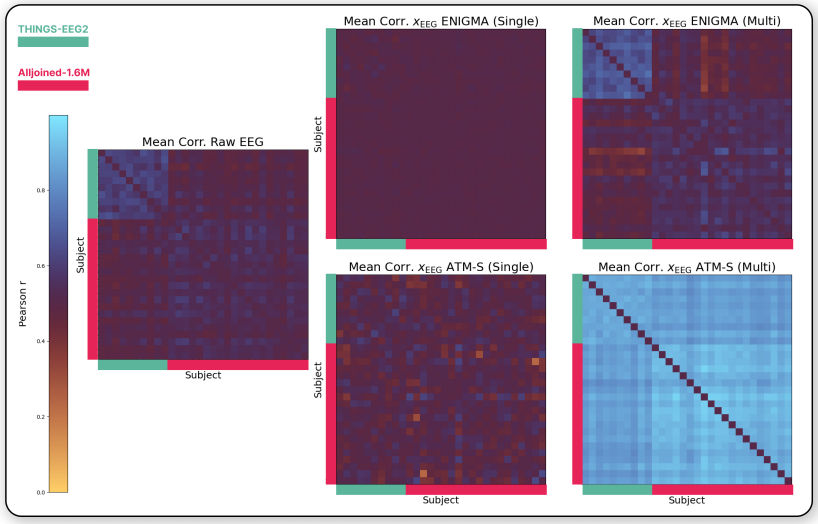

Figure 8: Subject-to-subject Pearson correlation matrices of predicted post-subject-wise-layer embeddings from different subjects in THINGS-EEG2 (green) and Alljoined-1.6M (red). Plots are displayed for **ENIGMA** and ATM-S in both single and multi-subject configurations.

correlations for within-dataset subjects and negative correlations for across-dataset subjects. In the case of multi-subject ATM-S, the attention layers have more or less ensured that all subjects correlate positively across categories but not too much as in that in CLIP latent space. This further strengthens our hypothesis that multi-subject ATM-S is overfitting in the CLIP latent space, hence explaining its poor performance in Table 1.

A.10 Behavioral Evaluation Experiments

To evaluate the quality of EEG-to-Image reconstruction models applied to our dataset, we conducted a behavioral experiment on 545 human raters online. For our experiment, we identified no risks to the human participants, and collected informed consent from all participants.

The experiment stimuli consists of image reconstruction sampled from the 30 subjects across THINGS-EEG2 and Alljoined-1.6M from all methods and cases in Table 1. The images were shuffled and 60 images presented to each subject. We use attention checks to identify whether human raters were paying attention to the task and the instructions and dropped 8 human raters who failed at least 2 out 8 attentions checks before analysis. An attention check presents the ground truth image as one of the candidate images and raters have to select the candidate ground truth image (as an image is most similar to itself) to pass.

Our subjects were recruited through the Prolific platform, with our experimental tasks hosted on Meadows. Each human rater was paid $1.25 for the completion of the experiment, and the median completion time was 5 minutes, resulting in an average payment rate of $15/hour. The code to reproduce our experiment can be found in our anonymized GitHub repository.

A.10.1 2AFC Identification Task

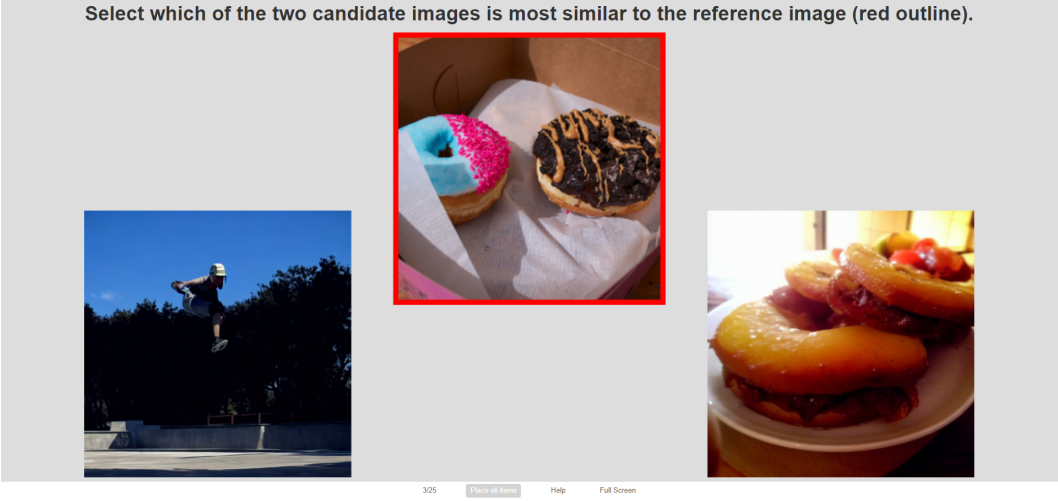

Figure 9: An example of the 2 alternative forced choice task used in our behavioral experiment performed by human raters.

Our experiment was a 2 alternative forced choice task (2AFC) facilitated by the "Match-To-Sample" task on the Meadows platform. An example of the first experiment can be seen in Figure 9. In this experiment, human raters were asked to select which of two candidate images was more similar to a reference image. The reference image provided is the ground truth image the subject either saw, and the 2 candidate images were the target reconstruction of the reference image, or a randomly selected reconstruction from an EEG recording corresponding to a different stimulus. The two candidate images were always sampled from the same reconstruction method and subject. This experiment was repeated for all reconstruction methods, model types, datasets, and subjects. With the results presented in Table 1, we establish a baseline for human-rated image identification accuracy of seen image reconstructions from EEG, as no other paper has conducted behavioral evaluations of EEG-to-Image reconstructions.