Alljoined-1.6M: A Million-Trial EEG-Image Dataset for Evaluating Affordable Brain-Computer Interfaces

Anonymous Author(s)

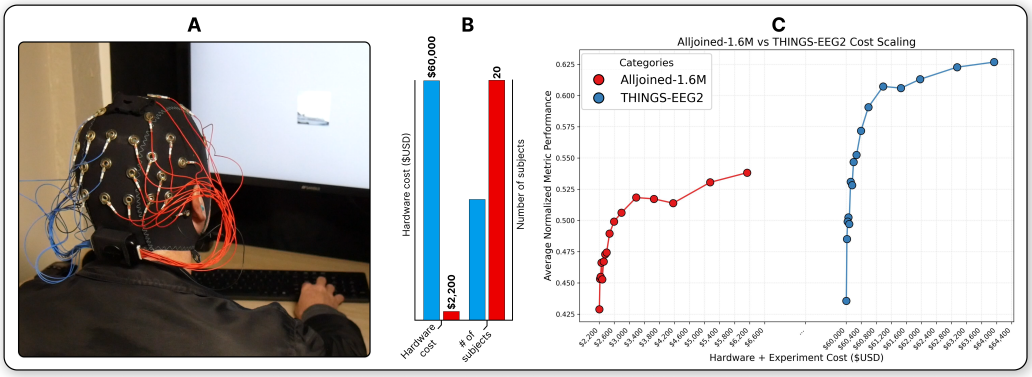

Figure 1: **A**: A picture of the collection setup for the Alljoined-1.6M dataset. **B**: A comparison between THINGS-EEG2 [17] and Alljoined-1.6M (ours) in terms of hardware cost and the number of subjects in the dataset. **C**: A bifurcated plot displaying decoding performance against collection cost. Costs include the purchase of the EEG headset and amplifier, and then a $50/hour collection cost to compensate participants and technicians.

Abstract

We present a new large-scale electroencephalography (EEG) dataset as part of the THINGS initiative, comprising over 1.6 million visual stimulus trials collected from 20 participants, and totaling more than twice the size of the current most popular benchmark dataset, THINGS-EEG2. Crucially, our data was recorded using a 32-channel consumer-grade wet electrode system costing ∼$2.2k - around 27x cheaper than research-grade EEG systems typically used in cognitive neuroscience labs. Our work is one of the first open-source, large-scale EEG resource designed to closely reflect the quality of hardware that is practical to deploy in real-world, downstream applications of brain-computer interfaces (BCIs). We aim to explore the specific question of whether deep neural network-based BCI research and semantic decoding methods can be effectively conducted with such affordable systems—filling an important gap in current literature that is extremely relevant for future research. In our analysis, we not only demonstrate that decoding of high-level semantic information from EEG of seen images is possible at consumer-grade hardware, but also that our data can facilitate effective EEG-to-Image reconstruction even despite significantly lower signal-to-noise ratios. In addition to traditional benchmarks, we also conduct analyses of EEG-to-Image models that demonstrate log-linear decoding performance with increasing data volume on our data, and discuss the trade-offs between hardware cost, signal fidelity, and the scale of data collection efforts in increasing the size and utility of currently available datasets. Our contributions aim to pave the way for large-scale, cost-effective EEG research with widely accessible equipment, and position our dataset as a unique resource for the democratization and development of effective deep neural models of visual cognition.

Preprint. Under review.

1 Introduction

Electroencephalography (EEG) is a widely used non-invasive method to measure human brain activity, with applications ranging from fundamental neuroscience to advanced brain-computer interfaces (BCI). A longstanding challenge in the development of BCIs is the trade-off between data quality and accessibility: high-density research-grade EEG systems provide high signal fidelity but are prohibitively expensive and resource-intensive, limiting the scope of data collection and the practicality of real-world applications. Consequently, most EEG studies rely on relatively small datasets, and real-world applications are rare. Despite these current limitations, recent advances in affordable EEG hardware and machine learning algorithms to decode brain data have begun to change this equation.

The cost of consumer-grade EEG systems has declined substantially, potentially lowering the barrier to BCI applications in medical and consumer sectors, as well as unlocking larger-scale neural data collection efforts. Devices such as the Emotiv EPOC and Flex series are available at a fraction of the cost of traditional research-grade systems, making them appealing for these scenarios. However, these affordable systems generally suffer from reduced signal-to-noise ratio (SNR) and other technical limitations [45, 38], and so research-grade EEG remains the standard for most current research efforts and datasets. Nevertheless, an important open question remains: *Can consumer-grade EEG systems facilitate shallow and deep neural decoding efforts?*

In this work, we aim to answer this question by combining the advantages of scale and accessibility. We introduce a new EEG dataset that is, to our knowledge, the largest of its kind, comprising more than twice the number of subjects as the previously largest human EEG object recognition dataset, THINGS-EEG2 [17]. Unlike THINGS-EEG2, which used a 64-channel research-grade EEG system, our dataset was collected with the Emotiv Flex 2, a 32-channel wireless headset retailing at roughly ∼$2.2k at the time of writing. Despite its lower cost, the Flex 2 can deliver full-scalp coverage and up to 256 Hz sampling (sub-4 ms temporal precision), enabling many experiments that historically required ∼$35 – $60k research-grade setups. Although the Flex 2 yields lower signal fidelity than research-grade systems, we demonstrate that it is nevertheless useful for facilitating downstream decoding tasks, such as semantic meta-category decoding, image retrieval, and EEG-to-Image reconstruction using state-of-the-art ML models.

Our contributions are three-fold:

1. **Dataset:** We release Alljoined-1.6M as part of the THINGS initiative, a large scale EEG dataset of visual perception containing 32-channel recordings from 20 participants, 4 sessions each, totaling 1.6 M trials across 16,740 unique images, all collected on affordable EEG hardware (Fig. 1A,B).

2. **Benchmarks and Decoding Analysis:** We provide extensive benchmarks and analysis of existing EEG-to-Image reconstruction, retrieval, and semantic decoding models on our dataset, setting a baseline for future research developing methods for decoding affordable EEG responses to visual stimuli.

3. **Scaling and Cost Analysis:** We conduct a detailed analysis of the within-subject scaling properties of our dataset, finding that, despite the lower SNR, decoding performance still increases log-linearly no signs of saturation, demonstrating that scaling is still an effective approach for improving decoding performance on consumer-grade EEG hardware. We also demonstrate (in Fig. 1C) the degree of financial investment necessary to obtain certain benchmarks for decoding performance.

Our findings demonstrate the growing potential of more cost-effective EEG headsets that mirror the hardware constraints of many real-world BCI deployments. By publicly releasing Alljoined-1.6M[1][2] and demonstrating its utility, we hope to democratize progress in EEG-based machine learning and lower the entry barrier for research groups with limited resources.

[1] https://huggingface.co/datasets/Alljoined/Alljoined-1.6M
[2] https://github.com/Alljoined/Alljoined-1.6M

2 Related Work

Low-Cost EEG Hardware. Consumer-grade EEG headsets from providers like Emotiv and OpenBCI have dramatically lowered the cost and logistical barriers to neural recording, enabling at-home and mobile experiments. Although these devices typically offer fewer channels and lower signal fidelity than clinical systems, a growing body of work demonstrates that even low-density EEG can yield meaningful biomarkers [4, 7, 29]. For example, Duvinage et al. [15] compared a 14-channel Emotiv EPOC headset against a 128-channel ANT medical-grade system on a P300 speller task. Although the Emotiv under-performed in single-trial classification, they concluded that the Emotiv could reliably support non-critical applications.

Deep Learning for Neural Decoding. Recent advances in deep learning—most notably transformer architectures [59], denoising diffusion probabilistic models [23], and contrastive language–image pretraining (CLIP) [43]—together with the availability of large-scale datasets of human brain activity—have catalyzed a new wave of neural decoding studies [41, 47, 48, 53, 54, 26–28, 32, 2, 16, 49]. These approaches learn rich spatiotemporal representations from neural recordings, enabling the decoding of high-level semantic information that eluded earlier methods of analysis such as the study of event-related potentials [35], EEG topographic analysis [37], and frequency-band metrics [42]. Such semantically informed decoding frameworks hold great promise for downstream brain–computer interface applications.

Scaling Laws in Neuroimaging. Recent research has demonstrated that scaling up neural data collection efforts can log-linearly improve modeling and decoding performance [5, 46], echoing trends observed in computer vision and natural language processing. Many advances in computer vision have been driven by massive, high-variance image corpora such as ImageNet, which contains over a million labeled images spanning a thousand object categories [13]. Inspired by this success, the neuroimaging community has pursued analogous scaling: MEG repositories like OMEGA and Cam-CAN amass hundreds of hours of data across dozens of participants [39, 57]; fMRI collections such as BOLD5000 and Generic Object Decoding sample tens of thousands of distinct stimuli [10, 25]; and the Natural Scenes Dataset comprises $\sim 30,000$ unique images viewed over 40+ 7T sessions at an estimated cost of \$450k [1]. Together, these efforts underscore a field-wide push for larger, more diverse datasets to deepen our understanding of brain representations and to power robust, generalizable models across modalities.

EEG Datasets and Benchmarks. Despite growing interest in deep learning for EEG across both decoding and representation learning [44, 31] domains, public EEG repositories remain overwhelmingly tailored to behavioral paradigms (e.g., motor imagery, sleep staging) or clinical biomarkers. Although there are available large-scale collections such as the TUH EEG Corpus with clinical recordings [40], and BCI competitions centered on specific paradigms [56], neither facilitates high-level semantic decoding. While initial efforts to collect EEG datasets in response to visual stimuli were revealed to have confounds in the block-design paradigm that allowed high-capacity models to exploit low-frequency drifts rather than genuine visual features [50, 33, 34], more recent contributions from the THINGS initiative have made strides in increasing the experimental rigor of such datasets: THINGS-EEG1 captured $22,248$ unique images across 50 subjects [20], and THINGS-EEG2 recorded $\sim 82,350$ trials per participant over $16,740$ stimuli with a 64-channel lab-grade system [17] and a shuffled block design with no overlap between training and testing image classes. While these datasets have been massively successful in enabling researchers to decode semantic content from EEG brain activity, they still rely on research-grade hardware, and often fall short of the scale and hardware accessibility constraints necessary to develop for robust, consumer-facing brain-computer interfaces. To address this gap, we release a multi-subject, million-trial EEG corpus collected entirely with a $\sim$\$2.2k, 32-channel consumer headset, along with code and benchmarks to enable semantic decoding research at scale.

3 Alljoined-1.6M

Hardware and Recording Setup. Alljoined-1.6M was collected using the Emotiv Flex 2 EEG system [3] with sintered silver/silver-chloride (Ag/AgCl) gel-based electrodes. The Flex 2 supports 32 EEG channels—we configured a montage covering primarily occipital regions associated with

[3] https://www.emotiv.com/products/flex-gel

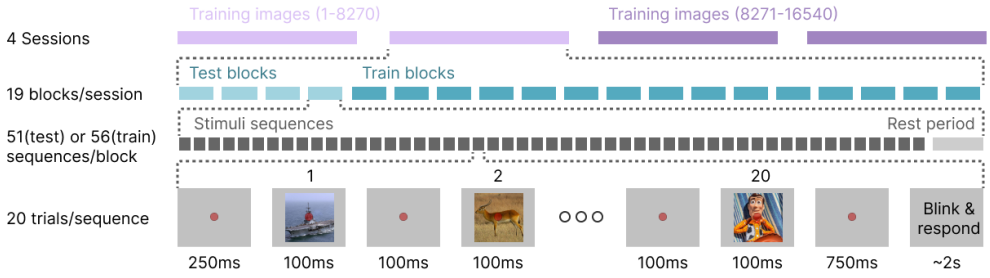

Figure 2: Experimental paradigm for Alljoined-1.6M. Details can be found in Section 3.

visual perception (detailed in Appendix A.1). The Flex 2 streams data wirelessly via Bluetooth 5.2, at a sampling rate of 256 Hz (sufficient to capture event-related potentials). Notably, the entire hardware setup (headset, sensors, software) cost only ∼$2.2k, in contrast to the research-grade 64-channel ActiChamp amplifier and cap used in THINGS-EEG2 which we estimated to cost ∼$60k. Participants wore the Flex 2 in a dark and quiet room, positioned 60cm from the screen displaying the stimulus images, with a viewing angle of 7°. Throughout each sequence (including blank trials), a small semi-transparent red fixation dot with a black border ($0.2° \times 0.2°$, 50% opacity) was present at the center of the stimulus. Stimuli were shown against a gray background with an RGB value of (127,127,127), and were presented with PsychoPy. Millisecond-accurate triggers delivered through the Emotiv API aligned the onset of a stimulus image with the timeline of EEG acquisition.

Participants. We recruited 20 healthy adult volunteers (ages 23-63, 15 male, 5 female) from local recruiting platforms in San Francisco, and filtered for participants who had high behavioral scores and high task engagement (from an original pool of 48 subjects). All had normal or corrected-to-normal vision, were provided written informed consent, and were compensated for four recording sessions scheduled on separate days. This approach follows the precedent of NSD [1] and THINGS-EEG2 [17] in obtaining repeated measurements for matched stimuli across multiple recording sessions. To ensure participant safety and to make potentially confounding variables explicit for downstream analyses, we employed two electronic questionnaires whose (anonymized) responses are distributed with the dataset. The screening form gathers stable participant traits—demographics, medical and neurological history, sensory status, prior neuro-imaging, neurodivergence, and data-use consent—to confirm eligibility and document potential confounds. The pre-session questionnaire records transient state variables before each visit (recent caffeine, alcohol or drug use, sleep, fatigue, and meal timing/content) so that session-specific physiological factors can be modeled or controlled. Together, these forms provide a transparent account of both stable (trait) and fluctuating (state) factors for every participant and session, enhancing the reproducibility and secondary-analysis value of the dataset.

Stimuli and Experimental Design. We used a rapid-serial visual-presentation (RSVP) paradigm paired with an orthogonal target-detection task to keep participants engaged (Fig. 2). Each trial consists of an image presented for 100ms, followed by a 100ms blank screen. All stimuli were drawn from the THINGS database [22], which contains 26,000 high-resolution photographs spanning 1,854 everyday object categories, and our experiment uses the same set of $16,740$ stimuli utilized in THINGS-EEG2 [17] to facilitate direct comparison. Importantly, there is no overlap in the images or image categories presented in the training and test partitions of the dataset, which helps reduce experimental confounds and downstream model overfitting. Each participant completed four recording sessions lasting ∼2 hours each, with each session comprising 19 RSVP blocks lasting ∼5 min each. The first 4 blocks of every session presented images from the 200 held-out test images shown in 51 RSVP sequences of 20 images per run, totaling $4 \times 51 \times 20 = 4,080$ trials for the test data. The remaining 15 blocks of each session presented the remaining $16,540$ training images, randomly split into two equal subsets that were displayed in sessions 1–2 and 3-4. Each test block consisted of 56 RSVP sequences of 20 images. Within a session, every image in the first subset was shown twice, giving four presentations across the two sessions. Images were randomized independently within each session, with the constraint that no image could repeat after fewer than two intervening items (i.e. an ABA or AA pattern was disallowed). To encourage vigilance without biasing perception toward any object category, our experiment also included attention check trials. At the end of each sequence, participants given up to 2s to press a key for whether they saw the catch trial of *Woody* appearing in ($\approx 6\%$ of images). Performance on these attention trials is described in

Appendix A.4. A video recording of the stimulus presentation, is also provided for reference on our Huggingface dataset.

Dataset Scale. Across the four sessions, each of the 20 participants completed $4 \times 20,880 = 83,520$ image trials, resulting in a total dataset size of $\sim$1.6M trials. Training images were repeated 4–5 times per participant, whereas each test image was shown 80 times, permitting a within-subject averaging procedure to be performed during inference to increase SNR. Total on-task recording time per participant was $\sim$8 hours, punctuated by brief breaks.

Data Processing and Format. Raw EEG was stored in standard `.edf` files and pre-processed with MNE-Python [19]. The Emotiv firmware first applies a dual 50/60 Hz notch filter, and continuous recordings were then epoched from -200 ms to 1000 ms relative to image onset. Synchronization mismatches in the Emotiv trigger stream led us to discard 0–0.6% of trials between all subjects, which was comparable to exclusion rates reported in earlier Emotiv evaluations [3, 61]. Epochs were baseline-corrected to the 200 ms pre-stimulus window and resampled to 250 Hz to match the format of the THINGS-EEG2 benchmark [17]. As a final preprocessing step, we performed multivariate noise normalization [21], to increase SNR, estimating the whitening matrix solely on the training partition to avoid training-test contamination.

Meta-Category Groupings. The original THINGS-EEG2 dataset, with its 1,854 fine-grained object categories, has been invaluable for benchmarking deep neural networks but poses challenges for simpler machine learning models and traditional ERP analyses, which often perform better on coarser distinctions facilitating insight into underlying brain activity. These simpler models are critical for leveraging low-cost EEG data in low-resource or real-time settings. Prior work has largely focused on the test set alone, without leveraging the train-test split structure to evaluate generalization across broader semantic boundaries from trained to unseen test images [17, 49]. To address this, we categorize all trials of our dataset into seven broad meta-category groupings—*Animals*, *Foods/Plants*, *Vehicles*, *Tools*, *Furniture/Household Items*, *Body Parts/Clothing*, and *Toys/Games/Musical Instruments*. For details on how these categories were created, see Appendix A.6. Our meta-categories are distributed consistently across training and test sets, enabling more interpretable classification and testing of generalization while preserving the dataset's fine-grained image metadata.

4 Analysis and Preliminary Results

We conducted a series of analyses to characterize the dataset and to evaluate the central question of our paper: can current consumer grade hardware facilitate meaningful downstream neural decoding research? We conduct a series of analyses across a wide range of downstream tasks, including semantic decoding, image retrieval, and EEG-to-Image reconstruction tasks.

ERP Analysis. To first demonstrate the reliability of the signal in this dataset, we visually inspected all subjects' session-wise and block-wise ERPs. We observed a pattern similar to the pattern shown in Fig. 3 which is typical of a 200ms interval RSVP experiment, with a $\sim$100ms peak (P1) and a $\sim$200ms trough N200 [35]. Then, we ran a non-parametric spatio-temporal cluster-based permutation test [36] to identify clusters of time points and electrodes where condition-specific ERPs diverged significantly across semantic categories. This was performed on averaged ERPs for each condition across trials, and differences were evaluated across subjects. Surprisingly, we found that 16 out of all 21 possible category comparisons yielded significant clusters ($p < 0.01$). The fact that we observe robust category-selective effects under these conditions—using low-cost, consumer-grade EEG hardware—underscores both the sensitivity of the paradigm and the potential suitability for such data for scalable, real-world BCI research. Cluster results are described in more detail in the Appendix A.7.

Pairwise Decoding. To examine when category information becomes linearly separable, we first adopted a pairwise Linear Discriminant Analysis (LDA) decoder. LDA is intentionally simple: it is fast enough to evaluate every post-stimulus sample, needs no hyper-parameter tuning, and yields a single weight vector per class pair, making the decision boundary transparent. Using the training split, we fit LDA models and computed time-resolved ROC-AUC on held-out trials. A cluster-based permutation test against the 0.50 chance level (multiple-comparison corrected) revealed significant clusters ($p < 0.01$) peaking around 100 ms, 220 ms, and 400 ms after stimulus onset (Fig. 4). Although the consumer-grade Emotiv headset in Alljoined-1.6M introduces more noise than the research-grade hardware used in THINGS-EEG2, the decoder still achieved robust

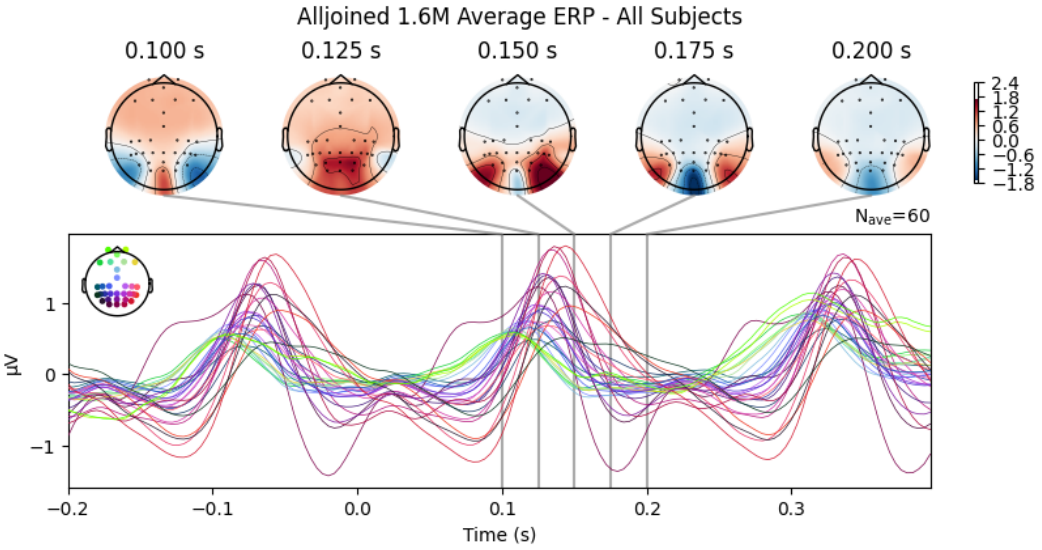

Figure 3: Average Event Related Potential (ERP) across all 20 subjects and all 4 sessions for a total of 1.6 million trials. Topographical maps show changes in visual cortex activity as expected for an RSVP experiment, with primary activity occurring in the occipital cortex for the duration of the image presentation. Three peaks show the ERP peaks of the 200ms interval between image display, essentially one peak per image shown.

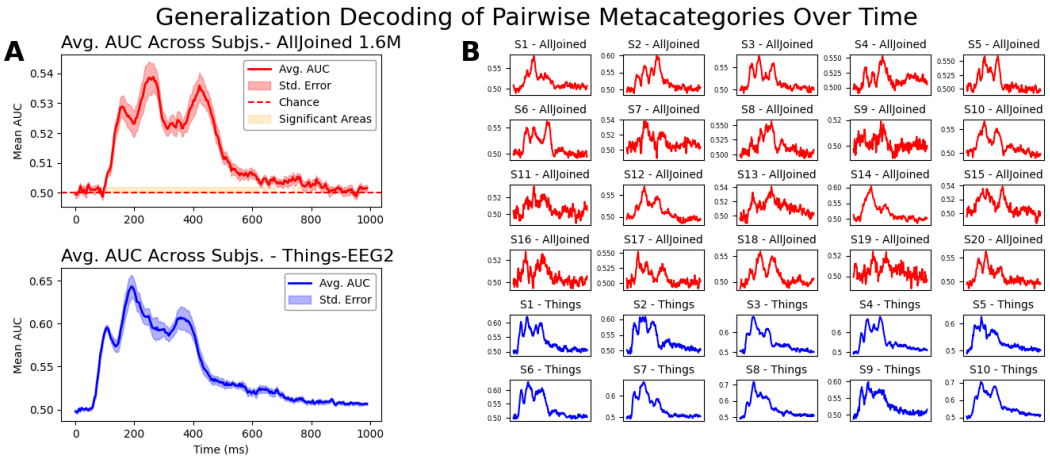

Figure 4: Average pair-wise decoding across meta-category combinations. Fig. 4A depicts the average decoding AUC scores across all 20 Alljoined-1.6M subjects (red) and across all 10 THINGS-EEG2 subjects (blue). Results show significant decoding compared to baseline. Fig. 4B Depicts the corresponding AUC performance scores for all individual subjects in both datasets.

above-chance performance, replicating the temporal structure reported with high-quality systems. Thus, a lightweight, interpretable linear model is sufficient to expose the key temporal dynamics of category-level signals in this dataset while providing a clear baseline for more complex approaches.

EEG-to-Image Reconstruction. One of the most promising areas of BCI research is the development of models trained to reconstruct seen images from human brain activity. The number of research efforts tackling the adjacent task of fMRI-to-Image reconstruction tasks has taken off recently [41, 47, 48, 53, 54, 26–28, 11]. However EEG-to-Image efforts have lagged behind, with the first large-scale EEG-to-Image datasets THINGS-EEG1 and THINGS-EEG2 released only in 2022 [20, 17]. For our analysis, we took all publicly available EEG-to-Image reconstruction methods (ENIGMA [2], ATM-S [32], and Perceptogram [16]) and reproduced their methods on our dataset. In line with research in fMRI-to-Image research [48], we also conducted a human behavioral experiment (n=545) to evaluate the identification accuracy of the reconstructions. Details on this behavioral

Figure 5: Qualitative comparison of EEG-to-Image reconstruction methods on 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.

experiment are provided in Appendix A.3. Image reconstructions from these methods can be seen in Fig. 5, and quantitative results in Table 1. We find that despite the high modeling difficulty of this task and the low SNR produced by the Emotiv hardware, several available EEG-to-Image reconstruction methods trained on Alljoined-1.6M produced reconstructions with quantitative scores comparable to those of THINGS-EE2 [2]. In our results we do notice that complex architectures like ATM-S [32] underperform on our data relative to simpler linear methods (Perceptogram [16], or more robust multi-subject models like ENIGMA [2]. We hope the release of Alljoined-1.6M will help spur further research translating existing approaches to the lower SNR data produced by the consumer-grade EEG setup in our study, as architecture clearly matters for bridging this gap.

Method	Low-Level		High-Level						Retrieval			Human Raters
	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓	Top-1 ↑	Top-5 ↑	Top-10 ↑	Ident. Acc. ↑
THINGS-EEG2												
ENIGMA	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	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	**0.247**	**0.431**	**85.46%**	88.03%	70.40%	71.98%	0.902	0.581	–	–	–	79.17%
Alljoined-1.6M												
ENIGMA	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	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	**0.094**	0.401	**67.36%**	**69.28%**	58.18%	59.94%	0.945	0.637	–	–	–	62.00%

Table 1: Quantitative comparison between reconstruction quality of available methods on the THINGS-EEG2 and Alljoined-1.6M datasets. PixCorr is the pixel-level correlation score. SSIM is the structural similarity index metric [60]. AlexNet(2) and AlexNet(5) are the 2-way comparisons (2WC) of layers 2 and 5 of AlexNet [30]. CLIP is the 2WC of the output layer of the CLIP ViT-L/14 Vision model [43]. Incep is the 2WC of the last pooling layer of InceptionV3 [52]. EffNet-B and SwAV are distance metrics gathered from EfficientNet-B13 [55] and SwAV-ResNet50 [9] models. Details on the human identification accuracy metric are provided in Appendix A.3. For EffNet-B and SwAV distances, lower is better. For all other metrics, higher is better. Bold indicates best performance, and underlines second-best performance. Additional details on the metrics used are in Appendix A.2.

Saliency Maps. To locate the spatiotemporal features that underpin our semantic predictions, we applied Integrated Gradients [51] with each meta-category's mean CLIP embedding as the target. For every category (animals, household items, foods/plants, tools) we averaged the training-image CLIP vectors, computed attributions over the full EEG tensor (channels × time), smoothed them with an 11-sample boxcar, and projected the result onto the 10-20 montage. All categories yield a pronounced attribution peak over occipital sensors at 160–300 ms post-stimulus (Fig. 6), pointing to early visual activity [12, 58].

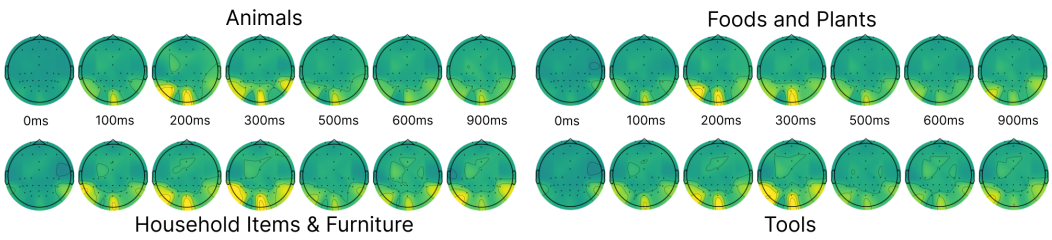

Figure 6: **Spatiotemporal saliency for ENIGMA.** Warm colors indicate electrodes whose activity diverges most strongly from the grand-average VEP and therefore pushes the model toward the CLIP centroid of that category. All four rows reveal a shared peak over occipital sensors around 160–300 ms (yellow), consistent with the early P1/N1 complex that dominates low-level visual processing.

Because our baseline is the grand-average EEG, the maps capture deviations from the canonical VEP that the network uses to match each category's CLIP centroid. The virtually identical occipital P1/N1 footprint across categories implies that the model relies almost entirely on low-level visual cues—an outcome we ascribe partly to the constraints of the RSVP paradigm [24].

Scaling Analysis. To contextualize Alljoined-1.6M within current scaling debates, we followed Banville et al.'s subsampling protocol [5] and trained ENIGMA [2], the leading EEG-to-Image model, on progressively larger subsets. Reconstruction quality was summarized by the normalized mean of the metrics in Table 1. Figure 7A plots ENIGMA's log–log learning curves for Alljoined-1.6M and, for comparison, the higher-SNR THINGS-EEG2. Performance increased almost linearly with log-trial count and showed no sign of saturating at the full dataset size. As expected, the consumer-grade recordings scaled less efficiently, underscoring the noise penalty of low-cost headsets. This limitation is also a strength: our dataset provides a realistic benchmark for methods that must cope with low-SNR data. The pattern mirrors findings across machine learning: more data reliably boosts accuracy - and the low price of consumer hardware should allow still larger datasets in the future, potentially offsetting the SNR gap through sheer quantity.

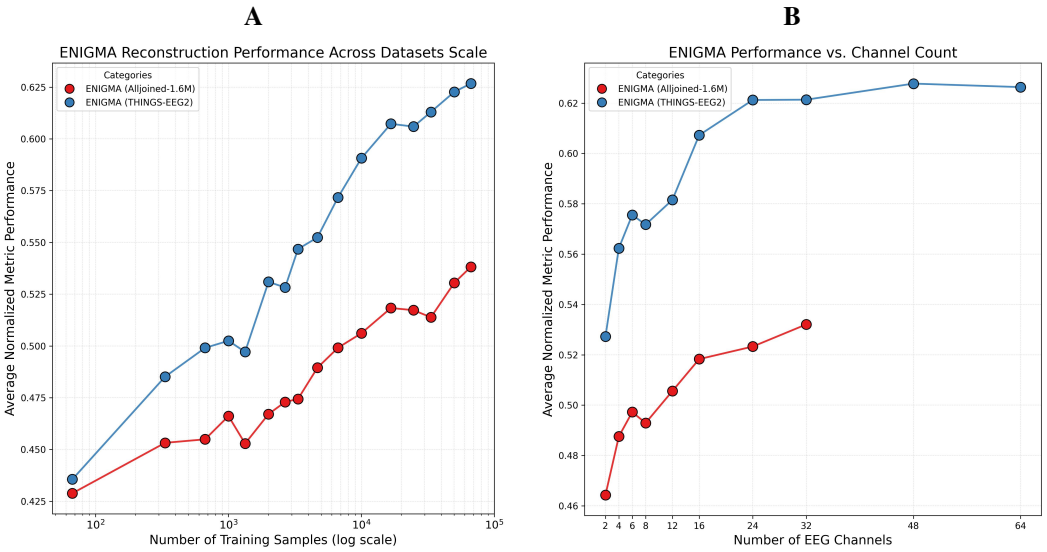

Figure 7: **(A)** Scaling analysis of model performance for Alljoined-1.6M and THINGS-EEG2. The number of training samples are plotted on a log-scale X-axis, and the normalized average of feature metrics presented in Table 1 is plotted on the Y-axis. **(B)** Channel count analysis of model performance 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. 7A.

Channel Count Analysis. One of the most obvious differences between the research grade ActiChamp amplifier used in the THINGS-EEG2 dataset and the Emotiv Flex 2 hardware used in our dataset is the number of channels (64 vs 32). We performed an analysis to evaluate how the

number of channels affects decoding performance using the ENIGMA model, and to explore whether this difference was a significant contributor to differences in performance. We sub-sampled both datasets to varying numbers of electrodes, while retaining a focus on covering occipital cortex with the electrodes selected. We find that while performance did drop with fewer channels, it is not the most significant factor accounting for the performance difference between the datasets, and that performance gains start to drop off after 24 channels, suggesting that future studies might still be able to achieve reasonable decoding performance with even fewer channels.

5 Discussion

We have introduced **Alljoined-1.6M**, the largest publicly available EEG dataset for visual cognition to date, and collected using a consumer-grade EEG headset. Our analyses provide encouraging evidence that EEG data collected on consumer-grade EEG hardware—such as the Emotiv Flex 2 utilized in our study—is still rich enough to train modern semantic decoding algorithms, a promising sign for the development of affordable brain-computer interfaces! We also find that scaling up data collection remains an effective way of increasing decoding performance even despite hardware limitations, opening doors to new large scale data collection efforts for affordable hardware. These findings have significant implications for the future of BCI research and cognitive neuroscience: large-scale EEG acquisition is now feasible for small labs, classrooms, and citizen-science projects. Alljoined-1.6M therefore serves as a benchmark for algorithmic progress, a blueprint for affordable data collection, and a concrete step toward democratizing neurotechnology.

Broader Impacts. Alljoined-1.6M highlights many considerations for the design of future datasets in brain decoding. While many efforts to collect neuroimaging datasets emphasize high channel counts or ultra-high-resolution signals, we believe that we need more datasets that are representative of real-world use cases. Our results (among others [5]) also point to the underexplored value of sheer volume, repetition, and participant diversity—factors that become significantly more tractable with affordable hardware. Alljoined-1.6M is a promising first step in shifting away from collecting a small amount of pristine signal, to instead optimizing for scale, signal diversity, and accessibility. We suggest that future datasets prioritize these axes, leveraging low-cost, high-throughput paradigms to explore larger-scale representational learning across subjects and tasks, much like large vision or language datasets [14, 8, 13] have done for deep learning. We envision a future where brain data collection is not bottle-necked by cost, and where massive EEG datasets fuel breakthroughs in understanding the brain and building BCI technologies to make a difference to people around the world.

Limitations and Future Work. Our dataset evaluates only one low-cost consumer-grade headset (Emotiv Flex 2). Testing other affordable devices, and guiding research-based product design in amplifiers and materials could further boost signal quality and sharpen the cost–performance frontier. We also observed roughly log-linear scaling; rigorously tracking accuracy as we grow from 10^6 to 10^7 trials, and exploring smarter sampling or augmentation strategies should help to further clarify these dynamics. Because the current corpus involves healthy adults in a controlled lab, future efforts should gather at-home, asynchronous recordings from diverse populations, transforming EEG collection into crowd-sourced neuroscience. It may also be exciting to merge multiple low-cost wearables or hybrid EEG + peripheral sensor arrays to narrow the gap to clinical-grade rigs. Methodologically, we observe the RSVP paradigm drops SNR at later latencies, so alternative task designs warrant exploration. Finally, our release offers a testbed for large-scale, multi-subject modeling: training a single network on the full corpus could yield generalizable neural representations transferable across users and downstream tasks, paralleling recent self-supervised EEG work [6].

Ethical Considerations. Efforts to collect and utilize neuroimaging datasets of human brain activity is rapidly growing in scale and capability. While this research promises clear downstream benefits in a variety of applications, we believe it is important to consider the ethical burden of gaining access to the internal cognitive states of individuals, and we recognize the potential for this technology to be misused. It is therefore important to begin developing an ethical framework for the application of brain-decoding devices and datasets that rigorously safeguards users data [18], and ensures that the technology is deployed transparently, responsibly, and for the benefit of humankind.

References

[1] 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.

[2] Anonymous. Enigma: A unified lightweight eeg-to-image model for multi-subject visual decoding. In Review, see Appendix B., 2025.

[3] 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.

[4] Nicholas A Badcock, Kathryn A Preece, Bianca de Wit, Katharine Glenn, Nora Fieder, Johnson Thie, and Genevieve McArthur. Validation of the emotiv epoc eeg system for research quality auditory event-related potentials in children. *PeerJ*, 3:e907, 2015.

[5] Hubert Banville, Yohann Benchetrit, Stéphane d'Ascoli, Jérémy Rapin, and Jean-Rémi King. Scaling laws for decoding images from brain activity. *arXiv preprint arXiv:2501.15322*, 2025.

[6] Hubert Banville, Yohann Benchetrit, Stéphane d'Ascoli, Jérémy Rapin, and Jean-Rémi King. Uncovering the structure of clinical eeg signals with self-supervised learning. *Journal of Neural Engineering*, 18(4):046020, 2021.

[7] Supriya Bhavnani, Dhanya Parameshwaran, Kamal Kant Sharma, Debarati Mukherjee, Gauri Divan, Vikram Patel, and Tara C Thiagarajan. The acceptability, feasibility, and utility of portable electroencephalography to study resting-state neurophysiology in rural communities. *Frontiers in human neuroscience*, 16:802764, 2022.

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[9] 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.

[10] Nadine Chang, John A. Pyles, Austin Marcus, Abhinav Gupta, Michael J. Tarr, and Elissa M. Aminoff. BOLD5000, a public fMRI dataset while viewing 5000 visual images. *Scientific Data*, 6(1):49, May 2019. Number: 1 Publisher: Nature Publishing Group.

[11] Zijiao Chen, Jonathan Xu, Jiaxin Qing, Ruilin Li, and Juan Helen Zhou. Structure-preserved image reconstruction from brain recordings, 2023.

[12] Radoslaw Martin Cichy, Dimitrios Pantazis, and Aude Oliva. Resolving human object recognition in space and time. *Nature neuroscience*, 17(3):455–462, 2014.

[13] 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*, pages 248–255. Ieee, 2009.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[15] Matthieu Duvinage, Thierry Castermans, Thierry Dutoit, Mathieu Petieau, Thomas Hoellinger, Caty De Saedeleer, K Seetharaman, and G Cheron. A p300-based quantitative comparison between the emotiv epoc headset and a medical eeg device. *Biomedical Engineering*, 765(1):2012–2764, 2012.

[16] 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).

[17] 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.

[18] 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.

[19] 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.

[20] Tijl Grootswagers, Ivy Zhou, Austin K. Robinson, Michael 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.

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

[22] Michael N. Hebart, Adam H. Dickter, Alexis Kidder, Anna Corriveau, Cody Van Wicklin, and Chris I. Baker. Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLOS ONE*, 14(10):e0223792, 2019.

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[24] Eric Lützow Holm, Diego Fernández Slezak, and Enzo Tagliazucchi. Contribution of image statistics and semantics in local vs. distributed eeg decoding of rapid serial visual presentation. *bioRxiv*, pages 2023–09, 2023.

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

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

[27] 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.

[28] 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. arXiv:2306.00927 [cs, q-bio].

[29] Michael T Knierim, Christian Zimny, Gabriel Ivucic, and Tobias Röddiger. Advancing wearable bci: Headphone eeg for cognitive load detection in lab and field. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(1):1–26, 2025.

[30] 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.

[31] Vernon J. Lawhern, Alex J. Solon, Nicholas R. Waytowich, Stacey M. Gordon, Christine P. Hung, and Brent J. Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018.

[32] 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.

[33] 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]. *arXiv preprint arXiv:1812.07697*, 2018.

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

[35] Steven J Luck. *An introduction to the event-related potential technique*. MIT press, 2014.

[36] Eric Maris and Robert Oostenveld. Nonparametric statistical testing of eeg-and meg-data. *Journal of neuroscience methods*, 164(1):177–190, 2007.

[37] Christoph M Michel and Thomas Koenig. Eeg microstates as a tool for studying the temporal dynamics of whole-brain neuronal networks: a review. *Neuroimage*, 180:577–593, 2018.

[38] Dmitry Mikhaylov, Muhammad Saeed, Mohamed Husain Alhosani, and Yasser F. Al Wahedi. Comparison of eeg signal spectral characteristics obtained with consumer-and research-grade devices. *Sensors*, 24(24):8108, 2024.

[39] Guiomar Niso, Christine Rogers, Jeremy T Moreau, Li-Yuan Chen, Cecile Madjar, Samir Das, Elizabeth Bock, François Tadel, Alan C Evans, Pierre Jolicoeur, et al. Omega: the open meg archive. *Neuroimage*, 124:1182–1187, 2016.

[40] Iyad Obeid and Joseph Picone. The temple university hospital eeg data corpus. *Frontiers in neuroscience*, 10:196, 2016.

[41] Furkan Ozcelik and Rufin VanRullen. Natural scene reconstruction from fMRI signals using generative latent diffusion. *Scientific Reports*, 13, 2023.

[42] Gert Pfurtscheller and FH Lopes Da Silva. Event-related eeg/meg synchronization and desynchronization: basic principles. *Clinical neurophysiology*, 110(11):1842–1857, 1999.

[43] 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.

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

[45] Joshua Sabio, Nikolas S Williams, Genevieve M McArthur, and Nicholas A Badcock. A scoping review on the use of consumer-grade eeg devices for research. *Plos one*, 19(3):e0291186, 2024.

[46] Motoshige Sato, Kenichi Tomeoka, Ilya Horiguchi, Kai Arulkumaran, Ryota Kanai, and Shuntaro Sasai. Scaling law in neural data: Non-invasive speech decoding with 175 hours of eeg data. *arXiv preprint arXiv:2407.07595*, 2024.

[47] 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.

[48] Paul Steven Scotti, Mihir Tripathy, Cesar Torrico, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A. Norman, and Tanishq Mathew Abraham. Mindeye2: Shared-subject models enable fMRI-to-image with 1 hour of data. In *ICLR 2024 Workshop on Representational Alignment*, 2024.

[49] 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.

[50] 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.

[51] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

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

[53] 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.

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

[55] 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.

[56] Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai J Miller, Gernot R Müller-Putz, et al. Review of the bci competition iv. *Frontiers in neuroscience*, 6:55, 2012.

[57] Jason R Taylor, Nitin Williams, Rhodri Cusack, Tibor Auer, Meredith A Shafto, Marie Dixon, Lorraine K Tyler, Richard N Henson, et al. The cambridge centre for ageing and neuroscience (cam-can) data repository: Structural and functional mri, meg, and cognitive data from a cross-sectional adult lifespan sample. *neuroimage*, 144:262–269, 2017.

[58] Simon Thorpe, Denis Fize, and Catherine Marlot. Speed of processing in the human visual system. *nature*, 381(6582):520–522, 1996.

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[60] 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. Conference Name: IEEE Transactions on Image Processing.

[61] 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.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, the downstream research goals and immediate research goals are laid out in the abstract and introduction, and are supported by our results throughout the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We provide a discussion of the current limitations of our research in Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: Our paper presents primarily empirical research and does not contain any proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Along with our submission, we release the dataset and code to reproduce our preprocessing steps. Code to reproduce many of our analyses can be found in the cited works for the open source models we utilized. A link to our dataset can be found in Section 1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Both the THINGS-EEG2 and Alljoined-1.6M are publicly available via their respective citations. Our source code for the ENIGMA model is linked in Section 1.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Details of the training procedures can be found in the citations for the open source methods utilized. Details of our data collection protocols are provided in Section 3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Yes, we provide a table of statistical significance measures in Appendix **??**.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: Significant computing resources were not used in our analysis. For the resources needed to train open source models utilized in our analysis, see the cited original works.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We make significant efforts to adhere to all ethical standards throughout our research.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Yes, both positive and negative societal consequences are discussed in Section 5.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our work poses no risks for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We own all data and assets released with this research paper, and use all existing datasets within their license restrictions.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Our dataset is well documented and released with all appropriate implementation details.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Section **??** discusses the protocols of our behavioral experiment used for reconstruction evaluations, as well as details about compensation. Section 3 discusses the protocols for our EEG data collection effort.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: While our research was not subject to an IRB, all participants in our experiment and dataset provided informed consent before participating, and no risks were posed to them in our research.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The meta-categories described in Section 3 were in part created by an LLM. For details on our process, see Appendix **??**

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.