A Appendix

A.1 Demographics

Table 2: *Participant Demographics by Category*

Variable	Category	Count	Percentage
Sex	Male	15	75
	Female	5	25
	Other	0	0
Age	18–25	3	15
	26–35	3	15
	36–45	6	30
	46–55	7	35
	55+	1	5
Ethnicity	White	4	20
	Black or African American	1	5
	Asian	3	15
	Hispanic/Latino	4	20
	Native Hawaiian or Pacific Islander	2	10
	American Indian or Alaska Native	1	5
	Multiracial	0	0
	Declined to answer	5	25
Handedness	Right	17	85
	Left	0	0
	Ambidextrous	3	15
Recruitment Platform	Craigslist	11	55
	Instawork	7	35
	Other	2	10

A.2 Data collection setup

The electrode positions selected for data collection followed the standard 10-20 format, and were focused on the occipital region and central line (Cz, Fp1, F7, F3, CP5, CP1, P1, P3, P5, P7, PO9, PO7, PO3, O1, O9, Pz, POz, Oz, O10, O2, PO4, PO8, PO10, P8, P6, P4, P2, CP2, CP6, F4, F8, Fp2). This corresponds to Layout 1 in 8). These were selected by running ablations of decoding performance on the Things-EEG2 dataset.

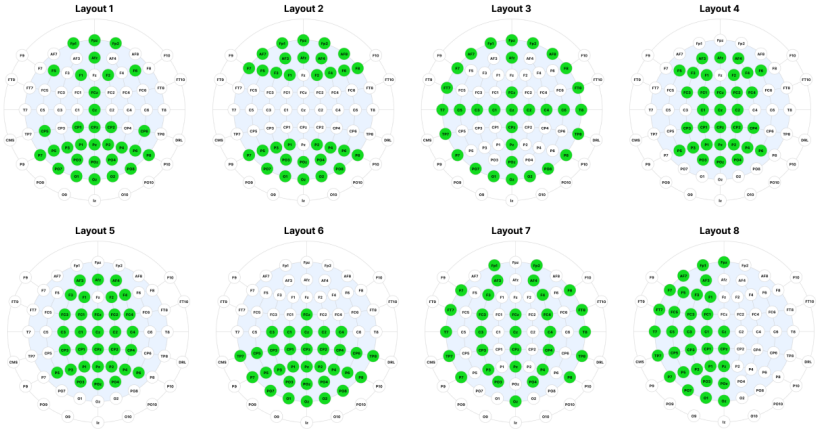

Figure 8: Various electrode positions and subsets from the 10-20 layout, compared in ablations on downstream decoding performance in Figure 9.

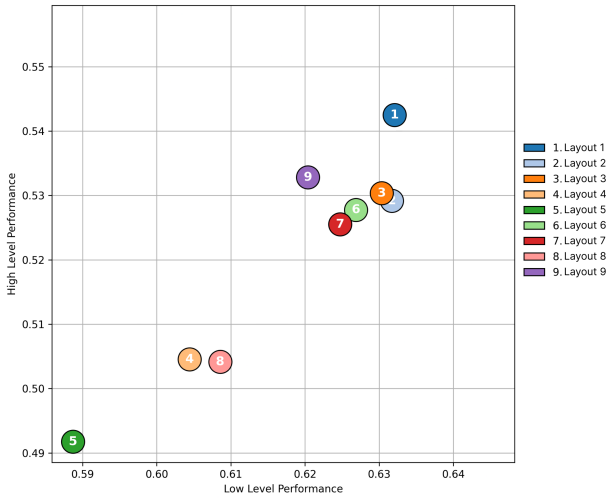

Figure 9: High and low level performance for image reconstruction on various 32 channel electrode subsets of THINGS-EEG2. Axes are the normalized average of the metrics in the high and low level metric categories displayed in Table 1.

A.3 Data Collection Details

Of the 48 participants initially recruited for this study, a total of 28 were subsequently excluded from the analyses for the following reasons. First, 18 participants were unable to commit to the three subsequent data collection sessions due to personal constraints, e.g., scheduling conflicts, illness, or travel. Second, despite our best efforts to coach and stabilize them, 6 participants produced data of insufficient quality as they were not able to sit still long enough for reliable EEG acquisition, resulting in excessive noise or artifacts. Finally, 4 participants were removed on the basis of behavioral and interpersonal factors: they were difficult or unpleasant to work with, or were unable to follow instructions for the experiment. The remaining 20 participants comprised the final dataset reported in this paper.

A.4 Behavioral Performance Analysis

During the experiment, we conducted regular attention checks in a manner very similar to the attention checks reported in the THINGS-EEG2 dataset [21]. Subjects were given the "odd-ball" task of looking for a picture of Woody from the Toy Story movies, and were asked at the end of each block whether an image of Woody appeared. We collected the accuracy of subjects on this detection task. However, since approximately 94% of the RSVP blocks did not contain the target (Woody), accuracy scores would be inflated by a strong response bias toward saying "no." That is, even random guessing would yield a high accuracy. To correct for this bias and better assess true sensitivity to the presence of the target, we computed the area under the receiver operating characteristic curve (AUC) for each participant and plotted them in Figure 10.

Subjects performed at a 88% AUC in the task, with a standard error of 1%. Three subjects performed close to chance rates, but all other subjects performed close to 100%. These results demonstrate that most subjects were attentive to the task.

The AUC reflects the probability that a randomly chosen trial with the target present is rated as more likely to contain the target than a randomly chosen target-absent trial. It is robust to response bias and provides a more balanced measure of detection performance in imbalanced-class settings. An AUC of 0.5 indicates chance-level performance, whereas 1.0 indicates perfect sensitivity.

We also note in our analysis that the three subjects that performed poorly on the attention task still yielded strong decoding performance on the visual decoding experiments, and so we keep these subjects in our dataset.

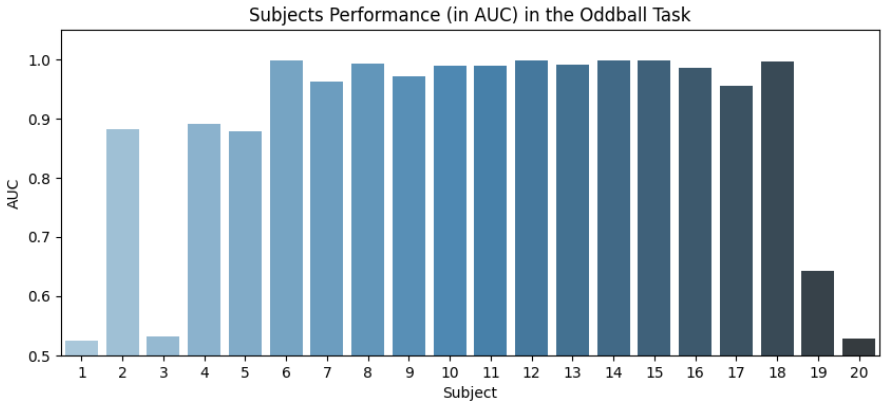

Figure 10: AUC values for each of the 20 participants based on their behavioral responses in the RSVP task. Higher AUC indicate better discrimination between target-present and target-absent trials, independent of response bias.

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 [66].
- AlexNet(2) and AlexNet(5) are the 2-way comparisons (2WC) of layers 2 and 5 of AlexNet [34].
- CLIP is the 2WC of the output layer of the CLIP ViT-L/14 Vision model [49].
- Incep is the 2WC of the last pooling layer of InceptionV3 [58].
- Eff and SwAV are distance metrics gathered from EfficientNet-B13 [61] and SwAV-ResNet50 [10] 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 [1], 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 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 3: 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 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 GitHub repository.

A.7.1 2AFC identification task

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 11. 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.

A.8 Meta-Categories

To create the meta-categories, we first used ChatGPT 4o and Gemini 2.5 Pro Preview to organize the original 1854 categories into groups. Note that the above GenAI tools often skipped words, misspelled them, miscategorized them, and sometimes hallucinated new words altogether. Therefore, we started with AI-generated categories and manually organized the images categories into meta-categories by also checking ambiguous or confusing categories (e.g., "mullet" refers to hair and not to fish).

Figure 11: An example of the 2 alternative forced choice task used in our behavioral experiment performed by human raters.

After visual inspection, we noticed that the test set did not contain any buildings and outdoor scenes, and excluded them from our analysis. We also noticed that the categories musical instruments, and toys and games were much smaller than the other categories, so we grouped them together in a "Fun/Entertainment" Category, although we note that these categories potentially evoke different brain responses.

A.9 Cluster Analyses

We ran all 21 contrasts between all pairs of metacategories, of which 16 had significant clusters. Most categories showed effects between 100 and 400ms, with the strongest and most consistent clusters emerging around 200ms post-stimulus at occipital electrodes—timing typically associated with early visual categorization components such as N170 and N200 [7, 18]. Additional clusters around 400ms may reflect later event related potential (ERP) components such as N2 [48] or N400 [35], or possibly contrastive or integrative responses to successive images. Given the 200ms inter-stimulus interval in our RSVP design, neural responses to consecutive images are likely to overlap in time, leading to temporally smeared effects across stimuli [20, 13]. This overlap might delay or attenuate the emergence of late components like P300 (which we did not observe in subjects' ERPs 3). Note that the data has been whitened, so the channels have no unit and do not look like traditional ERPs.

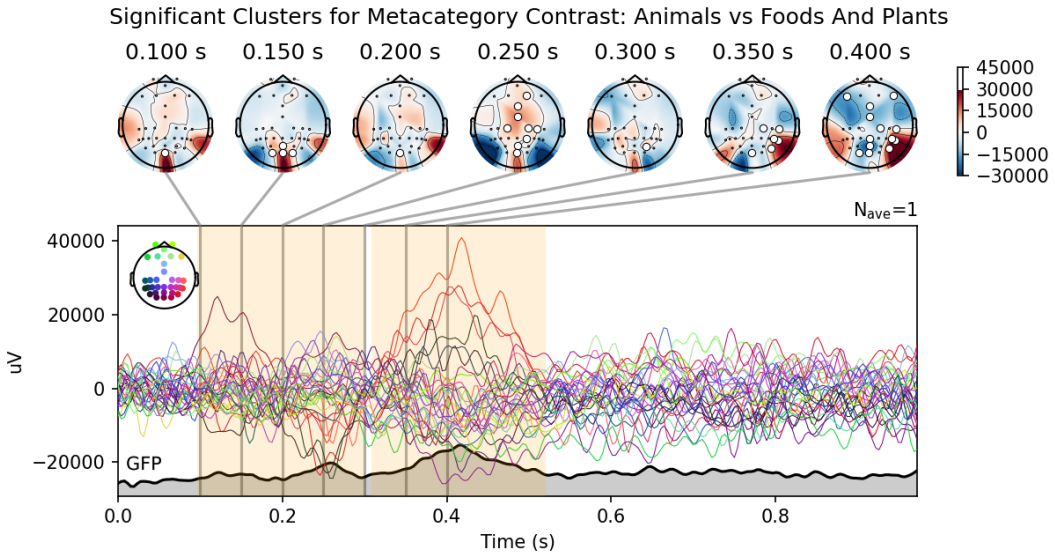

Figure 12: Cluster analyses results for the contrast between Animals and Foods/Plants. Yellow shaded area show times where difference between signals was significant. White dots in topographical maps represent significant electrodes which were significantly different for those time periods. Gray shaded area represents change in Global Field Power (GFP) over time.

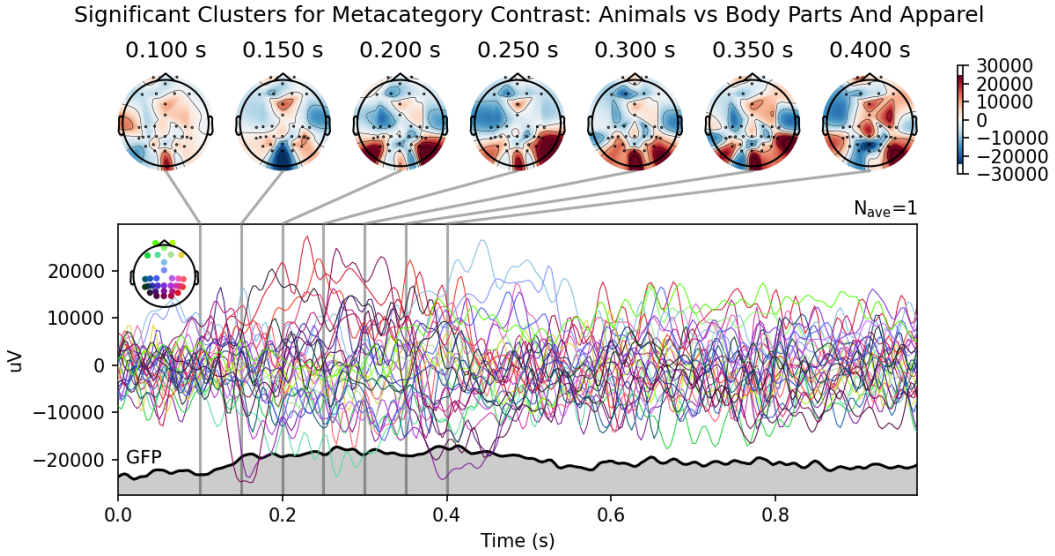

Figure 13: Cluster analyses results for the contrast between Animals and Body Parts/Apparel. Yellow shaded area show times where difference between signals was significant. White dots in topographical maps represent significant electrodes which were significantly different for those time periods. Gray shaded area represents change in Global Field Power (GFP) over time.

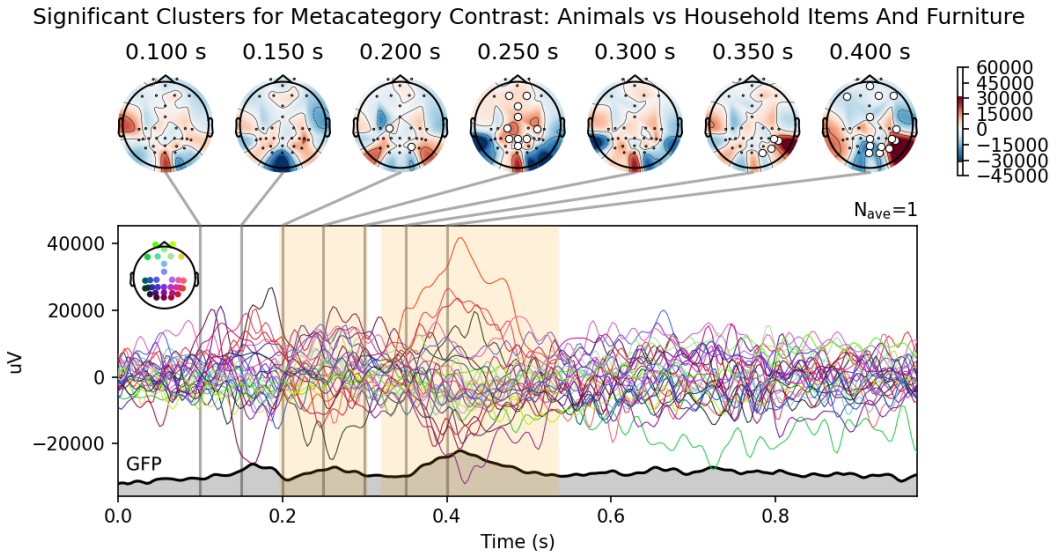

Figure 14: Cluster analyses results for the contrast between Animals and Household Items/Furniture. Yellow shaded area show times where difference between signals was significant. White dots in topographical maps represent significant electrodes which were significantly different for those time periods. Gray shaded area represents change in Global Field Power (GFP) over time.

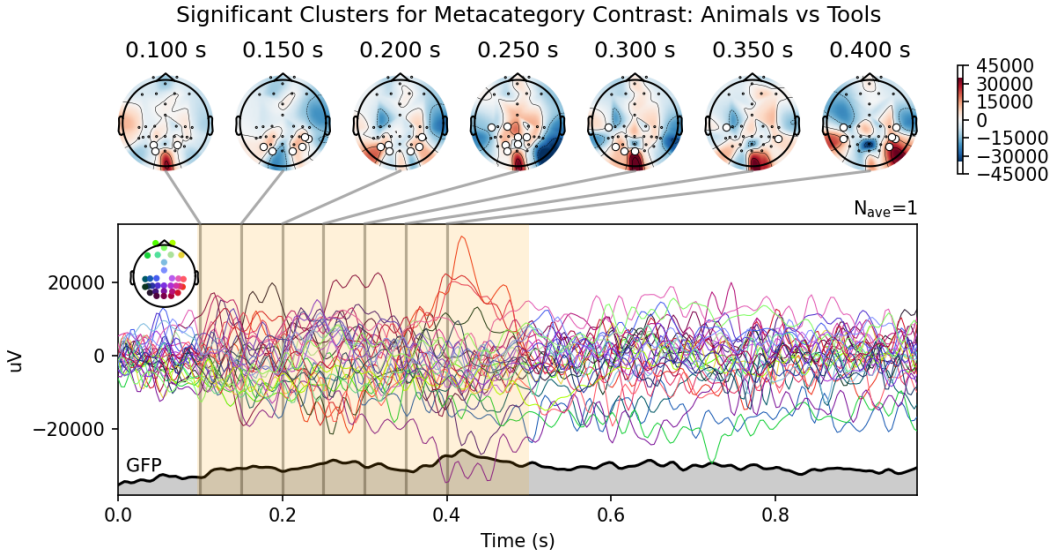

Figure 15: Cluster analyses results for the contrast between Animals and Tools. Yellow shaded area show times where difference between signals was significant. White dots in topographical maps represent significant electrodes which were significantly different for those time periods. Gray shaded area represents change in Global Field Power (GFP) over time.

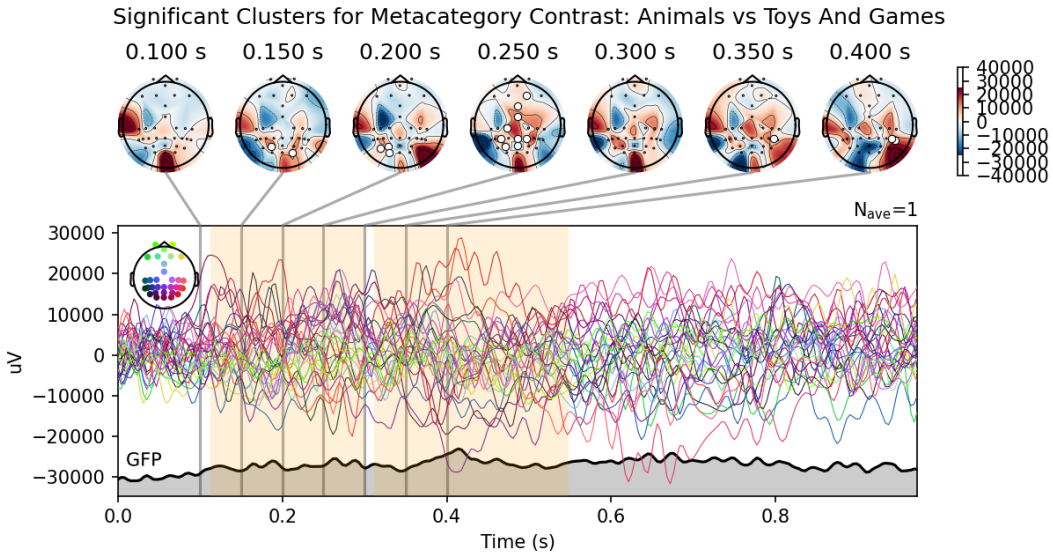

Figure 16: Cluster analyses results for the contrast between Animals and Toys/Games. Yellow shaded area show times where difference between signals was significant. White dots in topographical maps represent significant electrodes which were significantly different for those time periods. Gray shaded area represents change in Global Field Power (GFP) over time.

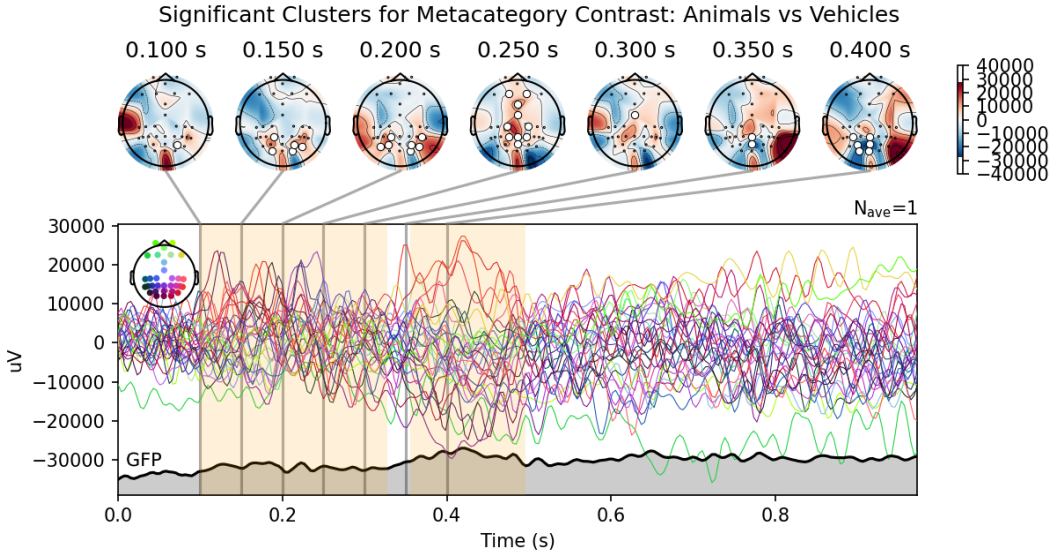

Figure 17: Cluster analyses results for the contrast between Animals and Vehicles. Yellow shaded area show times where difference between signals was significant. White dots in topographical maps represent significant electrodes which were significantly different for those time periods. Gray shaded area represents change in Global Field Power (GFP) over time.

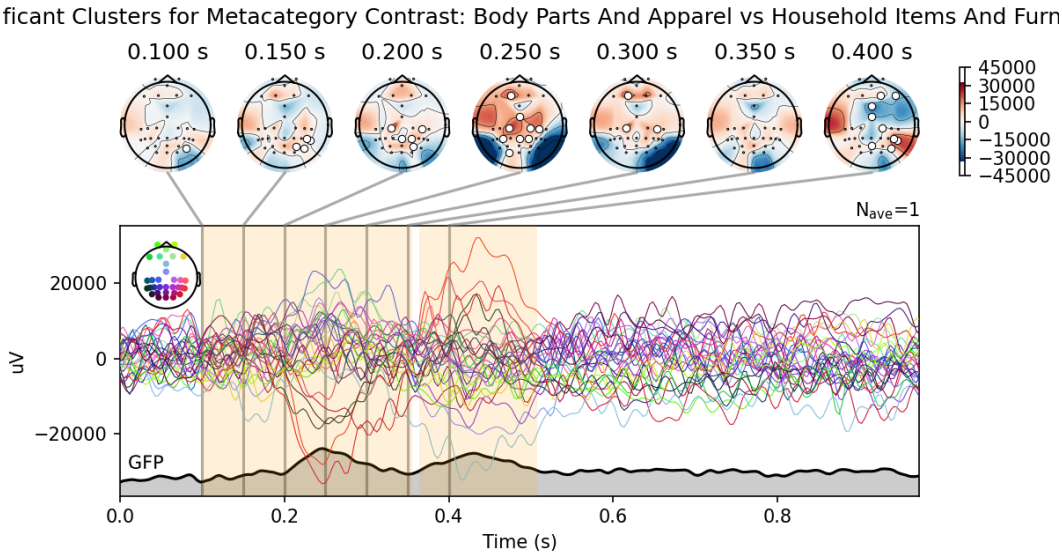

Figure 18: Cluster analyses results for the contrast between Body Parts/Apparel and Household Items/Furniture. Yellow shaded area show times where difference between signals was significant. White dots in topographical maps represent significant electrodes which were significantly different for those time periods. Gray shaded area represents change in Global Field Power (GFP) over time.

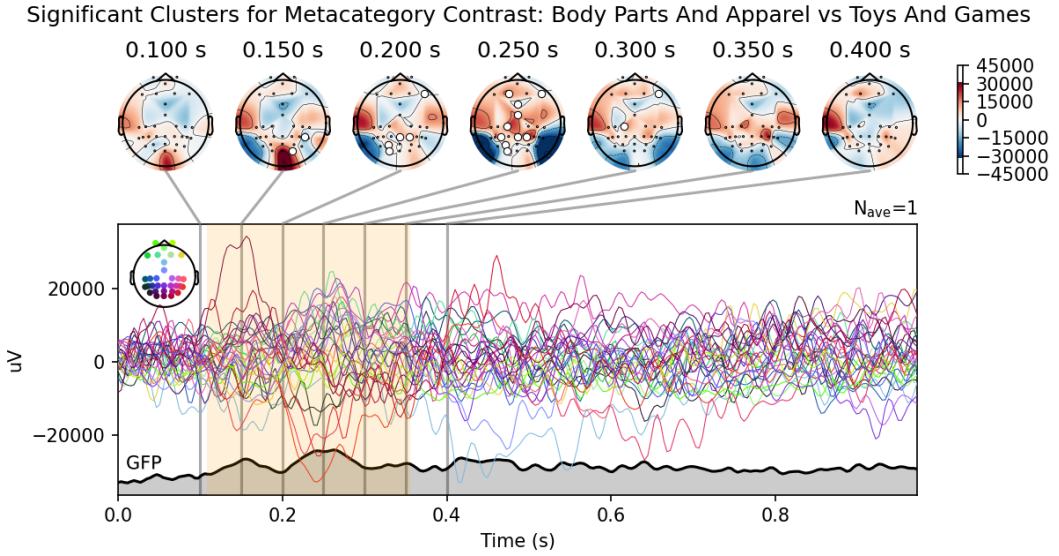

Figure 19: Cluster analyses results for the contrast between Body Parts/Apparel and Toys/Games. Yellow shaded area show times where difference between signals was significant. White dots in topographical maps represent significant electrodes which were significantly different for those time periods. Gray shaded area represents change in Global Field Power (GFP) over time.

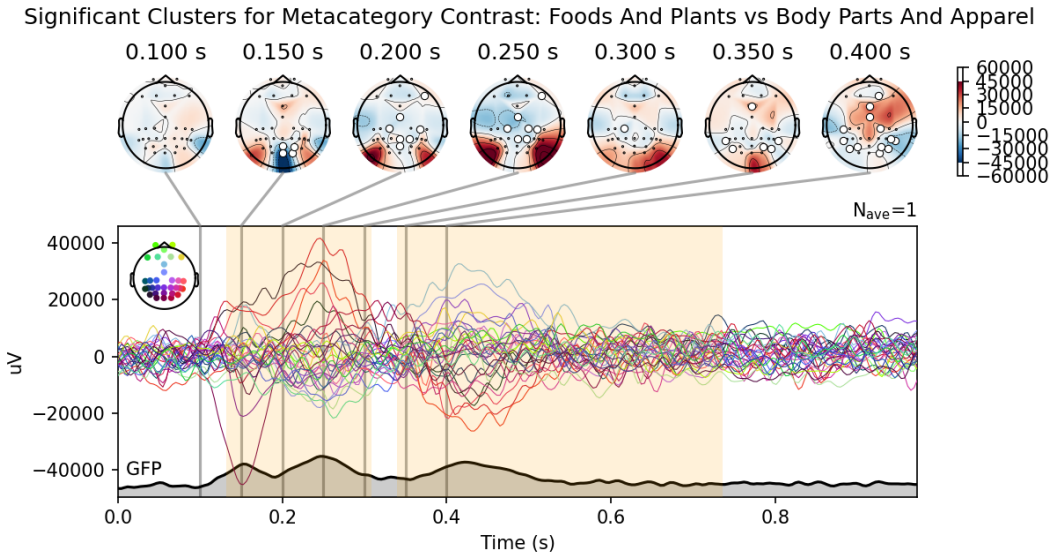

Figure 20: Cluster analyses results Foods/Plants and Body Parts/Apparel. Yellow shaded area show times where difference between signals was significant. White dots in topographical maps represent significant electrodes which were significantly different for those time periods. Gray shaded area represents change in Global Field Power (GFP) over time.

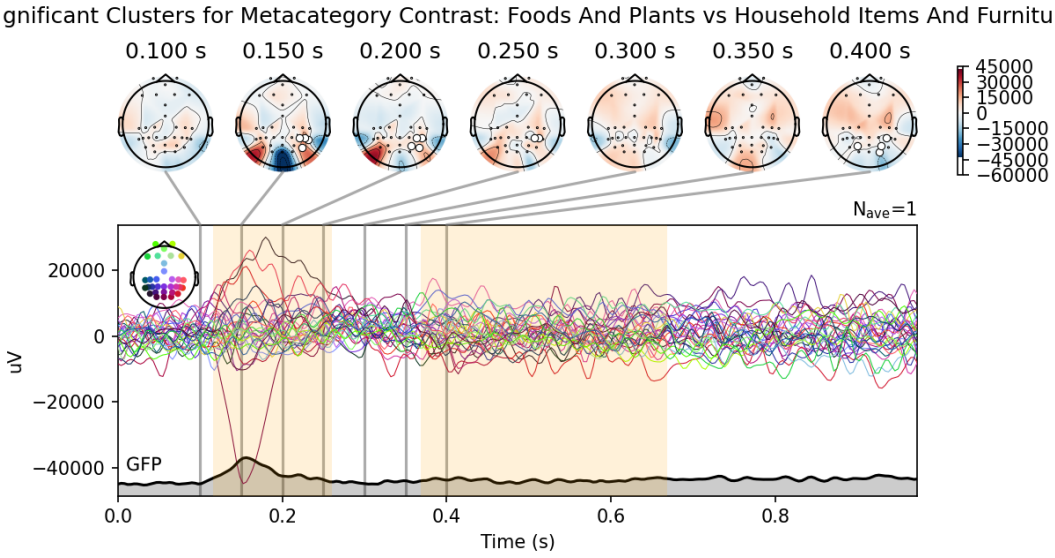

Figure 21: Cluster analyses results for the contrast between Foods/Plants and Household Items/Furniture. Yellow shaded area show times where difference between signals was significant. White dots in topographical maps represent significant electrodes which were significantly different for those time periods. Gray shaded area represents change in Global Field Power (GFP) over time.

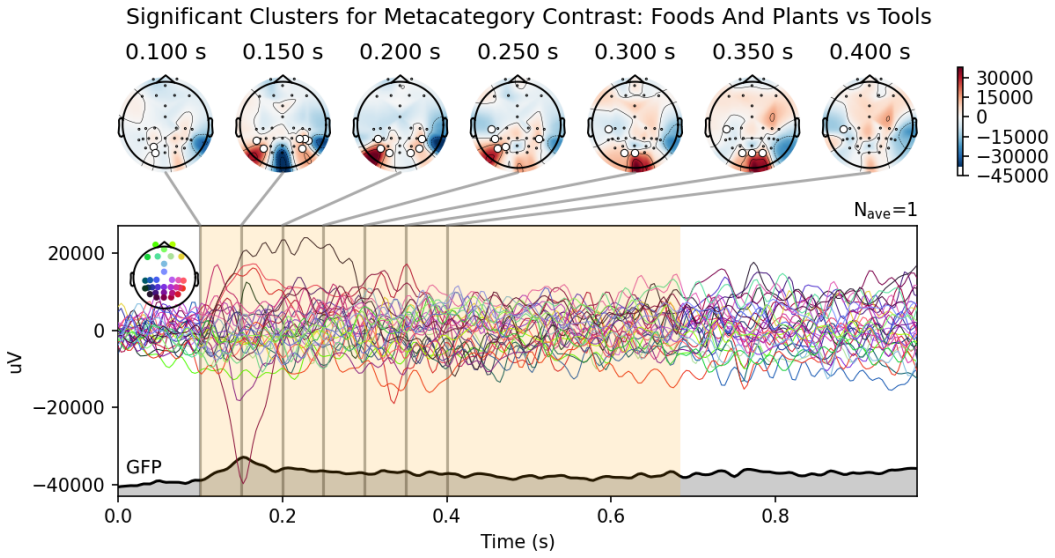

Figure 22: Cluster analyses results for the contrast between Foods/Plants and Tools. Yellow shaded area show times where difference between signals was significant. White dots in topographical maps represent significant electrodes which were significantly different for those time periods. Gray shaded area represents change in Global Field Power (GFP) over time.

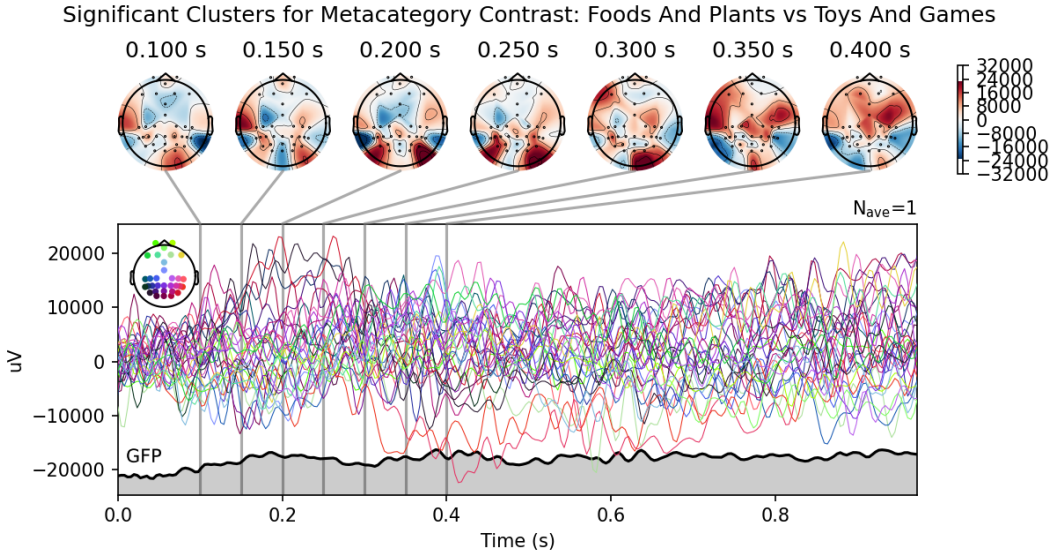

Figure 23: Cluster analyses results for the contrast between Foods/Plants and Toys/Games. Yellow shaded area show times where difference between signals was significant. White dots in topographical maps represent significant electrodes which were significantly different for those time periods. Gray shaded area represents change in Global Field Power (GFP) over time.

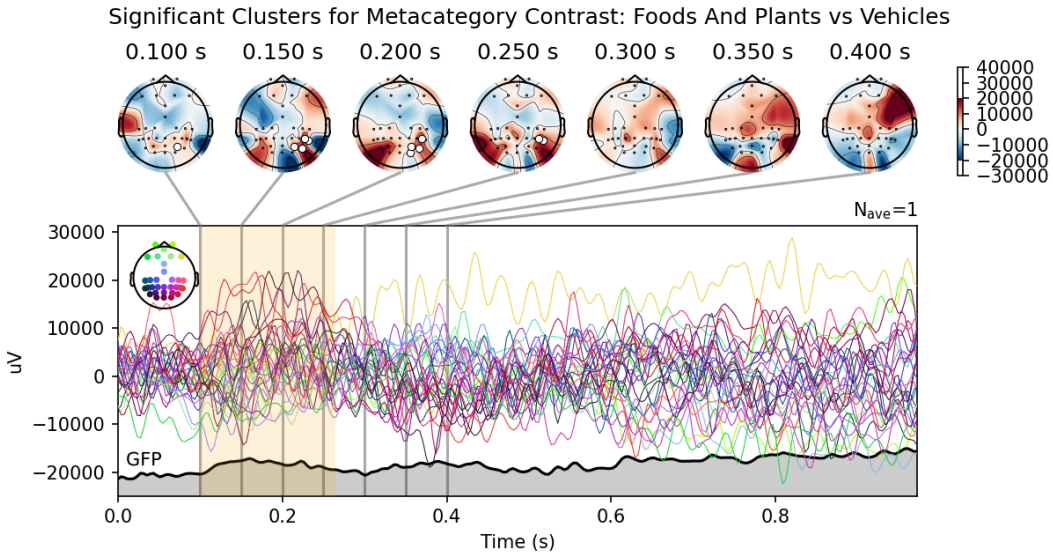

Figure 24: Cluster analyses results for the contrast between Foods/Plants and Vehicles. Yellow shaded area show times where difference between signals was significant. White dots in topographical maps represent significant electrodes which were significantly different for those time periods. Gray shaded area represents change in Global Field Power (GFP) over time.

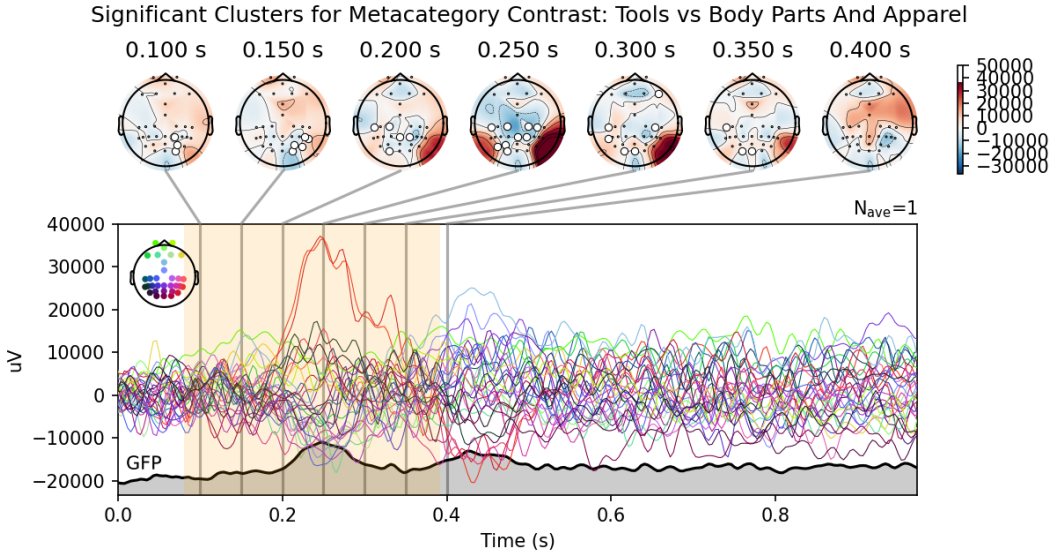

Figure 25: Cluster analyses results for the contrast between Tools and Body Parts/Apparel. Yellow shaded area show times where difference between signals was significant. White dots in topographical maps represent significant electrodes which were significantly different for those time periods. Gray shaded area represents change in Global Field Power (GFP) over time.

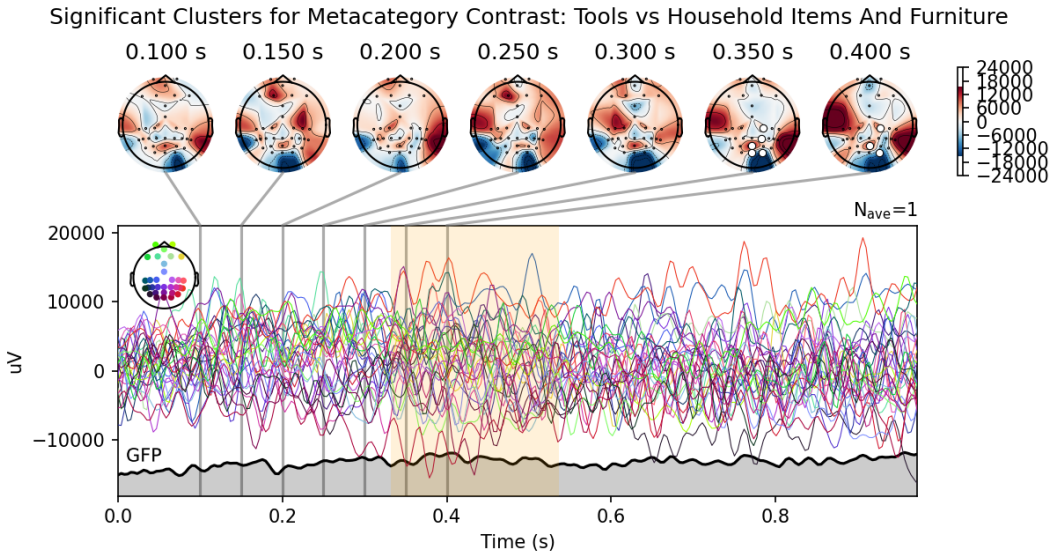

Figure 26: Cluster analyses results for the contrast between Tools and Household Items/Furniture. Yellow shaded area show times where difference between signals was significant. White dots in topographical maps represent significant electrodes which were significantly different for those time periods. Gray shaded area represents change in Global Field Power (GFP) over time.

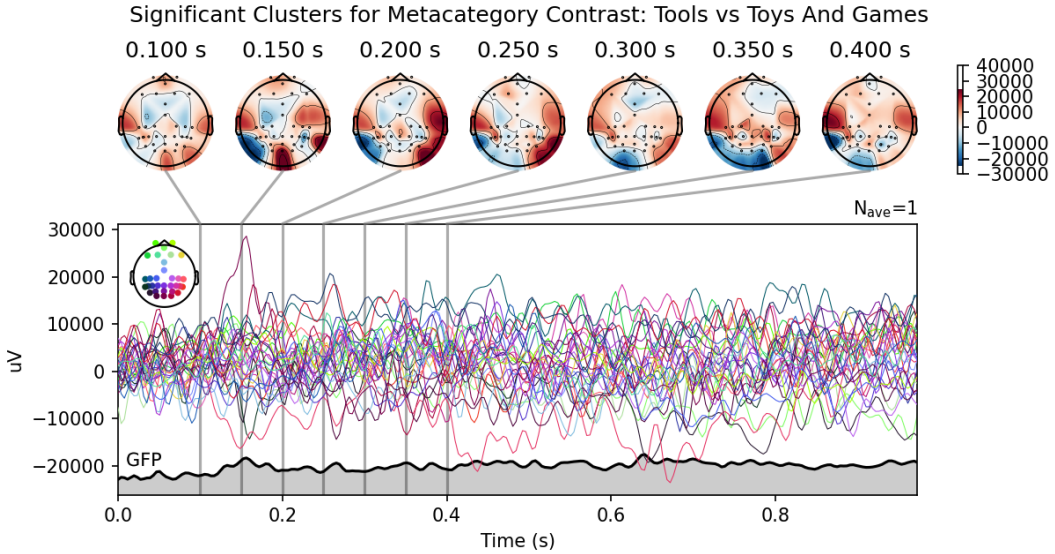

Figure 27: Cluster analyses results for the contrast between Tools and Toys/Games. Yellow shaded area show times where difference between signals was significant. White dots in topographical maps represent significant electrodes which were significantly different for those time periods. Gray shaded area represents change in Global Field Power (GFP) over time.

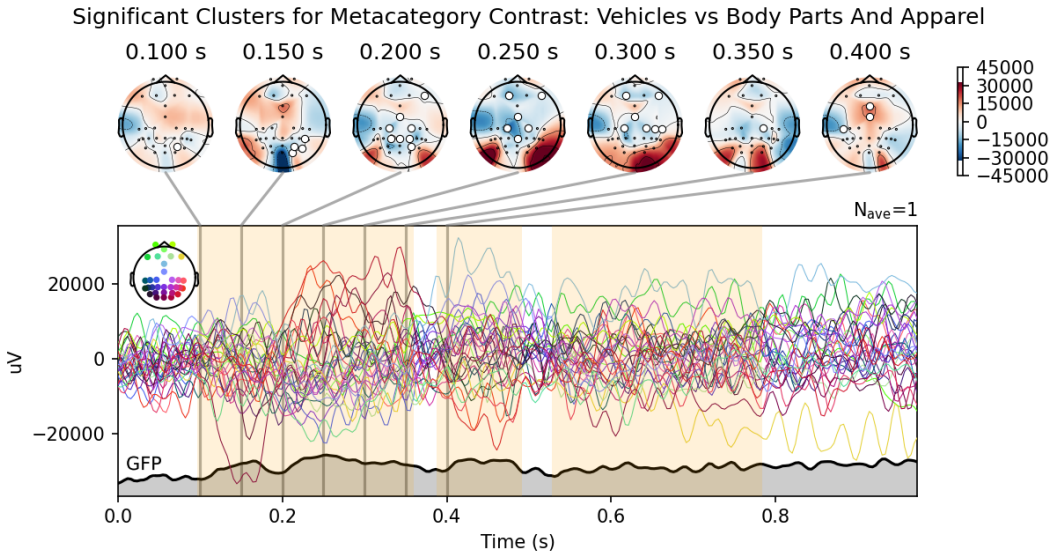

Figure 28: Cluster analyses results for the contrast between Vehicles and Body Parts/Apparel. Yellow shaded area show times where difference between signals was significant. White dots in topographical maps represent significant electrodes which were significantly different for those time periods. Gray shaded area represents change in Global Field Power (GFP) over time.

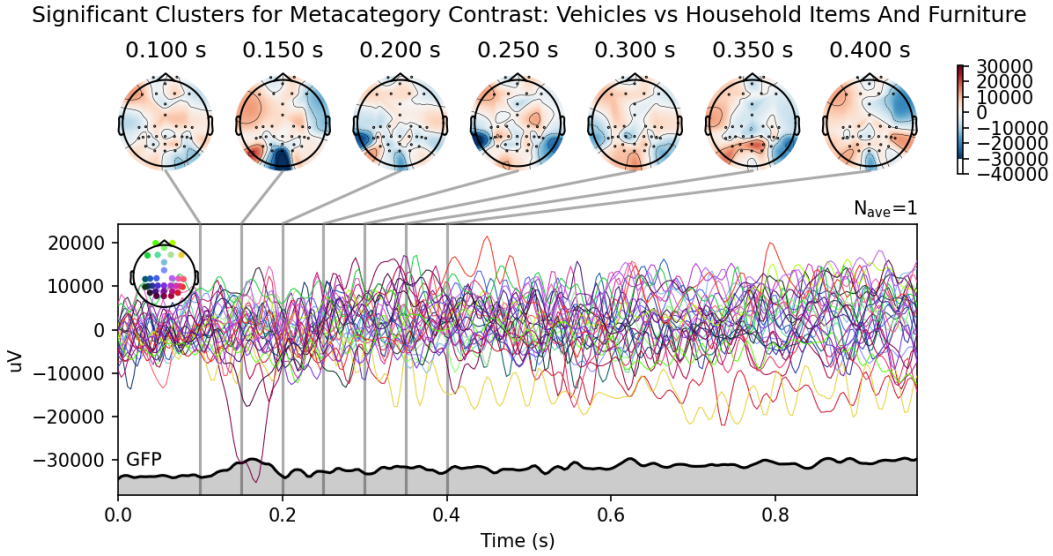

Figure 29: Cluster analyses results for the contrast between Vehicles and Household Items/Furniture. Yellow shaded area show times where difference between signals was significant. White dots in topographical maps represent significant electrodes which were significantly different for those time periods. Gray shaded area represents change in Global Field Power (GFP) over time.

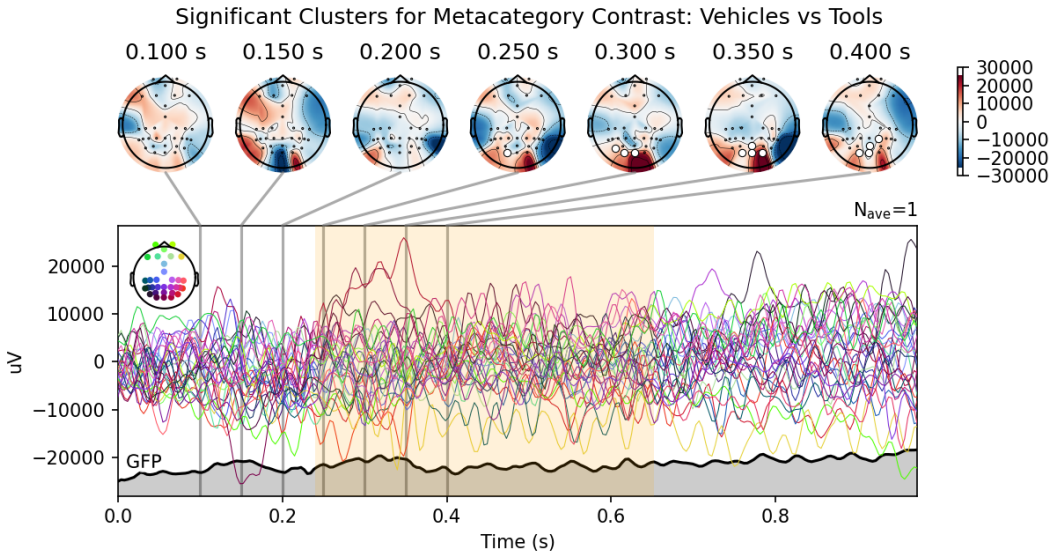

Figure 30: Cluster analyses results for the contrast between Vehicles and Tools. Yellow shaded area show times where difference between signals was significant. White dots in topographical maps represent significant electrodes which were significantly different for those time periods. Gray shaded area represents change in Global Field Power (GFP) over time.

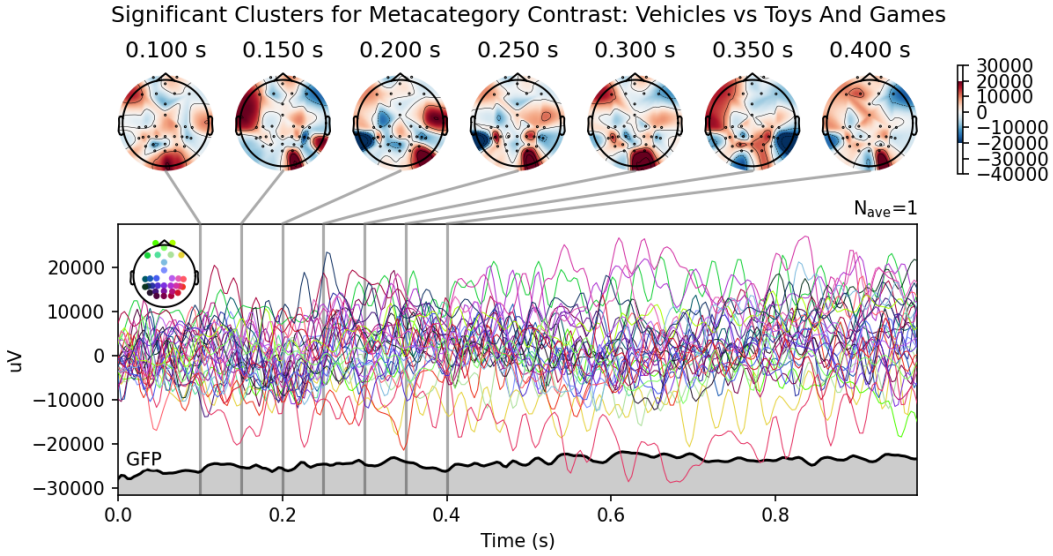

Figure 31: Cluster analyses results for the contrast between Vehicles and Toys/Games. Yellow shaded area show times where difference between signals was significant. White dots in topographical maps represent significant electrodes which were significantly different for those time periods. Gray shaded area represents change in Global Field Power (GFP) over time.

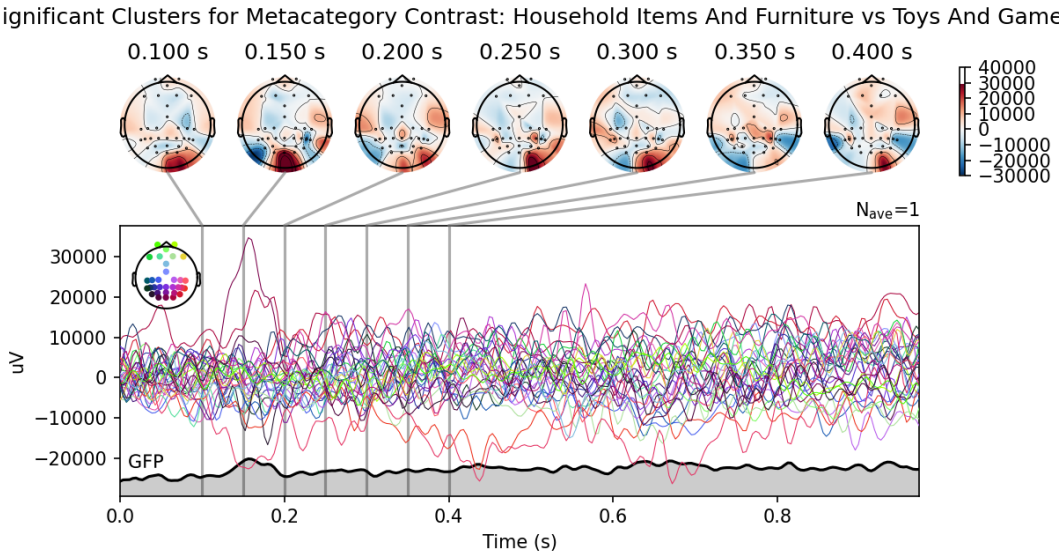

Figure 32: Cluster analyses results for the contrast between Household Items/Furniture and Toys/Games. Yellow shaded area show times where difference between signals was significant. White dots in topographical maps represent significant electrodes which were significantly different for those time periods. Gray shaded area represents change in Global Field Power (GFP) over time.

B Concurrent Submission

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

Anonymous Author(s)

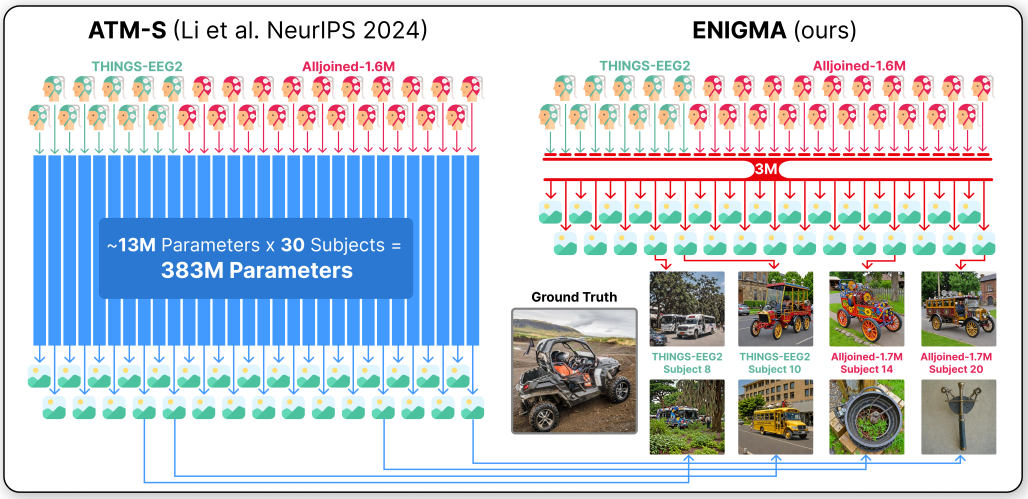

Figure 1: **ENIGMA** (ours) vs ATM-S [1] comparison of model size, methodology, and reconstructions of a seen image from EEG brain activity across subjects from the THINGS-EEG2 [2] (green cap) and Alljoined-1.6M [3] (red cap) datasets.

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 $\sim$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 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

Preprint. Under review.

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 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], 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.

Existing EEG-to-Image decoding research spans a wide range of architectural approaches: Fei et al. [7] demonstrates a simple linear mapping from EEG to an expressive CLIP (Contrastive Language-Image Pretraining [8]) image embedding space, and when combined with a pre-trained diffusion model suffices to produce recognizable images, while Li et al. [1] proposes a highly complex architecture (ATM-S) utilizing a transformer-based brain encoder and a two-stage generation process utilizing the diffusion prior introduced in Scotti et al. [9]. 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.

The recent release of the Alljoined-1.6M dataset—designed for evaluating the effectiveness of these methods on affordable hardware—provides a tool for refining EEG-to-Image models to be robust to drops in hardware quality and incorporate large numbers of subjects. We believe that these and the ability to run locally on accessible computing resources are fundamental requirements for the development of practical BCI applications, and warrant significantly more attention in current research. Existing benchmarks on Alljoined-1.6M [3] demonstrate 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 [10, 7] 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. **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, a standard in the adjacent field of fMRI-to-Image research.
2. Our model is unified across subjects (and datasets), resulting in a $\sim$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 when compared to single-subject modeling approaches.
3. We conduct detailed analyses of our method and previous approaches on EEG-to-Image decoding: we investigate 1) which aspects of their architectures are most effective across differences in hardware quality/multi-subject use cases, and 2) the alignment consistency of the decoded embedding space in single and multi-subject configurations on both research-grade and consumer-grade EEG hardware setups.

Our work demonstrates that a subject-unified model can better harness the information in EEG recordings to reconstruct visual percepts with unprecedented clarity. We believe **ENIGMA** will serve

[1] https://anonymous.4open.science/r/ENIGMA-9C46/

as a foundation for future BCI research and applications, bringing us closer to practical real-time "mind-to-image" translation.

2 Related Work

EEG-based Visual Decoding. 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. [11], many contained confounds that made it difficult to decode true semantic content [12], 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. [13] 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 [10, 1] employing contrastive learning between image and EEG features have achieved significant gains in zero-shot object classification. Alljoined-1.6M, The latest release from the THINGS initiative, 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.

Diffusion Models for Brain Decoding. The advent of generative diffusion models has revolutionized neural decoding for adjacent scanning modalities like fMRI [14, 9, 6, 15–19]. Takagi and Nishimoto [15] demonstrated some of the first high-resolution image reconstructions from fMRI by mapping fMRI activity into the latent space of a diffusion model. Ozcelik and VanRullen [14] were the first to show that latent diffusion can reconstruct natural scenes from fMRI with high semantic fidelity, and defined a set of evaluation metrics combining low-level (pixel-wise) and high-level (feature-based) similarity measures that have become standard [14, 9, 6, 15–19]. While fMRI enables finer-grained reconstructions due to its high spatial resolution, EEG's superior temporal resolution and portability makes it more suited for real-time applications, despite its lower signal fidelity.

In the space of EEG-to-Image reconstructino methods, Li et al. [1] introduced a specialized EEG encoder called the Adaptive Thinking Mapper (ATM-S), which uses a two-stage decoding approach comprising a transformer, a CNN, an MLP, and a diffusion prior before using the decoded CLIP vector to generate an image reconstruction using a diffusion module. One downside of this approach is the complexity of the ATM-S architecture, which requires careful tuning training of many intricate architectural components, and multiple sequential training stages. It was also shown with the release of the Alljoined-1.6M dataset that this degree of complexity results in brittle performance on lower grade EEG hardware [3]. While the architecture does support multi-subject training through a learned subject embedding in the transformer stage, training ATM-S on multiple subjects produces a substantial performance drop, so for most analyses in this work we use the model in its single-subject configuration.

Inspired by Ozcelik and VanRullen [14], Perceptogram [7] utilizes a linear transform to map EEG recordings to a CLIP embedding space, and generates images directly from the predicted embeddings using a diffusion model. While reconstructions still contain less detail than fMRI-based reconstruction, the approach of [7] demonstrates the significant power of robust linear models in produced recognizable images from low SNR brain activity patterns, and achieved state-of-the-art performance on the THINGS-EEG2 dataset. The authors also observed that EEG reconstructions preserve certain categories (e.g., animals, food) better than others, and linked them to existing EEG signatures.

3 ENIGMA

3.1 Methodology

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: 1) a set of subject-specific linear encoder layers, 2) a spatio-temporal convolutional neural network to learn the semantic information encoded in the spatial and temporal dimensions of the input signal, 3) an MLP projector to the ViT-H/14 CLIP [8] embedding space, and 4) an image reconstruction module, i.e., SDXL.

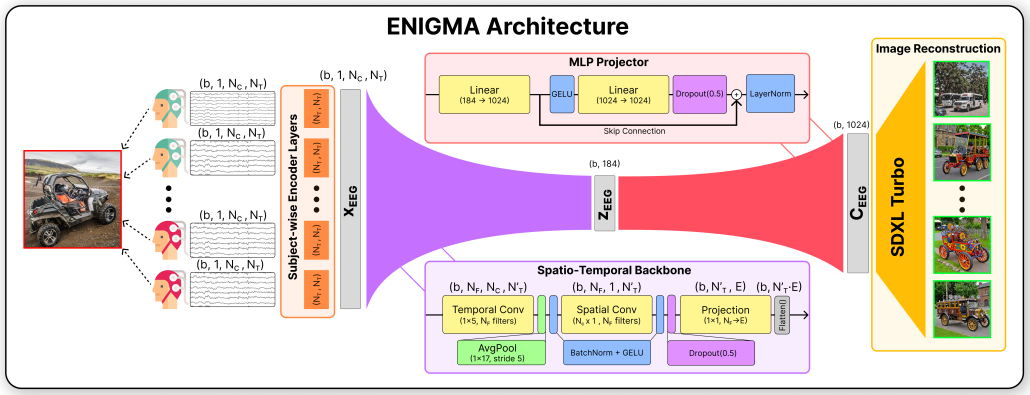

Figure 2: During training, brain activity from each subject is passed through its own subject-specific encoder layer, after which the embeddings x_{EEG} from all subjects are passed through a shared pathway of spatio-temporal convolutions, producing an intermediate latent vector z_{EEG}. This latent is passed through a fully connected MLP projection layer to produce the output c_{EEG} vector. Details of these procedures are provided in Section 3.3.

3.2 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 data preparation and preprocessing steps.

3.3 Architecture

The following components of our **ENIGMA** architecture are depicted in Figure 2.

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. We treat 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:

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

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:

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

We apply dropout [20] ($p = 0.5$) 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 After obtaining the z_{EEG} feature vector, we use a projection head to map it to the final CLIP ViT-H/14 latent dimension $D = 1024$. The head 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 To generate images from the EEG embedding c_{EEG}, we leverage Stable Diffusion XL Turbo (SDXL) [21], and its associated CLIP ViT-H/14 image-prompt adapter (IP-Adapter) [22]. 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 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 [23, 8]. 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 $\lambda = 0.5$

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 4.3, which shows that the inclusion of the diffusion prior negatively impacts reconstruction performance for our architecture.

Our overall loss function is

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

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, $\sim$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.

4 Results

We evaluate **ENIGMA** against the only available EEG-to-Image baselines, Perceptogram [7] and ATM-S [1], and report results on two of the most prominent benchmarks: THINGS-EEG2 [2] and Alljoined-1.6M [3].

Figure 3 presents a set of the best reconstructed images from EEG, comparing our method with available baselines on both available datasets. These examples illustrate typical outcomes: our

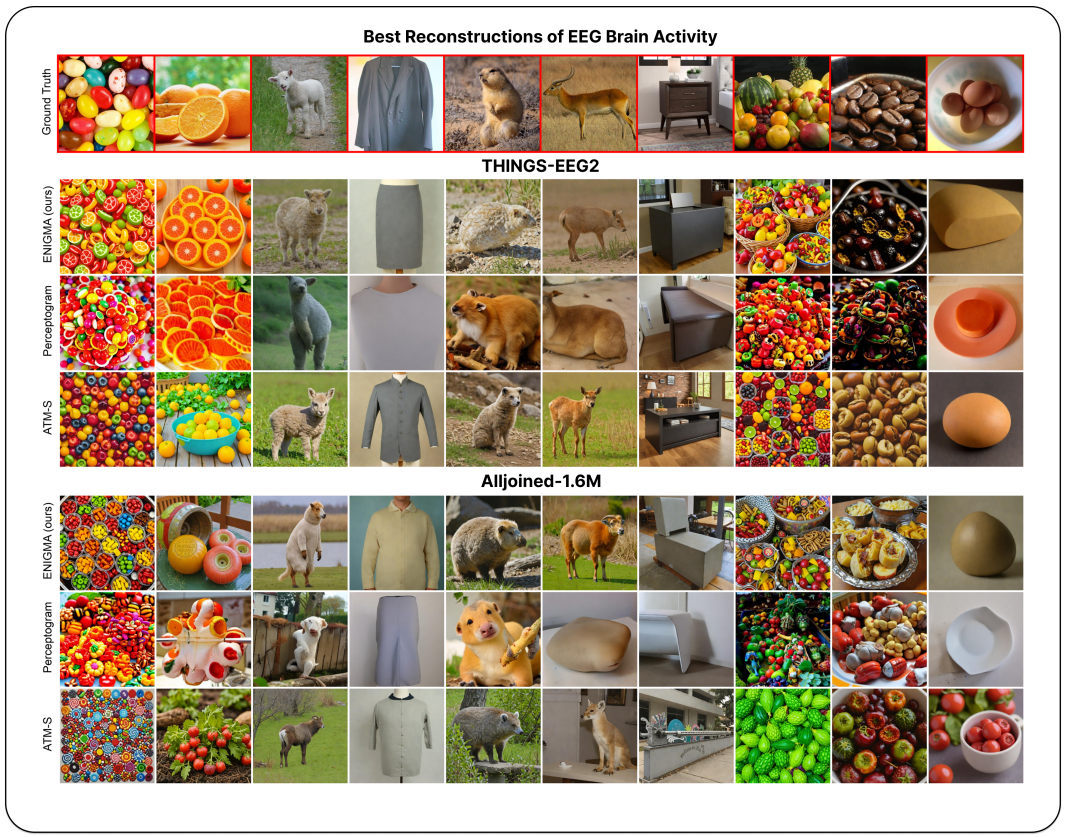

Figure 3: Qualitative comparison of reconstruction methods on seen stimuli from THINGS-EEG2 and Alljoined-1.6M. Reconstructions selected are the outputs sampled from each method and stimulus with the highest scores on all of the image feature metrics in Table 1.

reconstructions generally capture the correct high-level object, e.g., oranges, sheep, furniture, etc. Perceptogram's are usually blurrier and sometimes miss the object entirely, e.g., producing a vague shape or significant visual distortions. The ATM-S images are categorically similar, but are often less visually specific to the object being decoded.

4.1 Quantitative Evaluations

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 4.2. 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.2.

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 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 [24] or quality [25]. Thus, we also provide human identification accuracy scores in the "Human Raters" section, discussed further in Section 4.2.

4.2 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.5.

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.

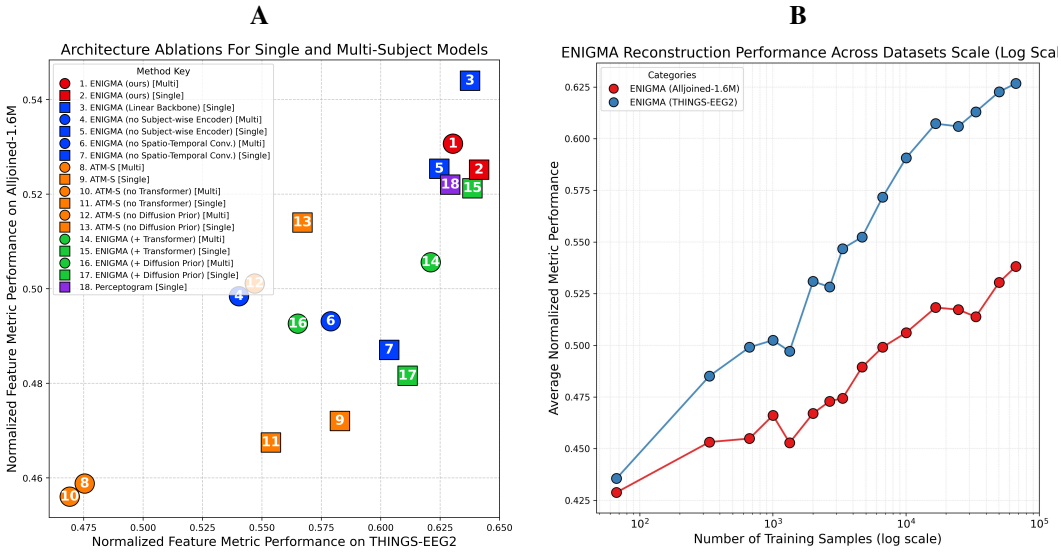

Figure 4: **(A)** Ablation analyses: model variants (numbered icons) in single (square) and multi-subject (circle) configurations under each ablation type (color) are assessed via the normalized average of all feature metrics (Table 1), with THINGS-EEG2 performance on the x-axis and Alljoined-1.6M performance 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.

4.3 Ablation Study

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] to remain a highly effective way to decode semantic information from brain activity, although we note that linear models do not provide any of the parameter efficient 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.

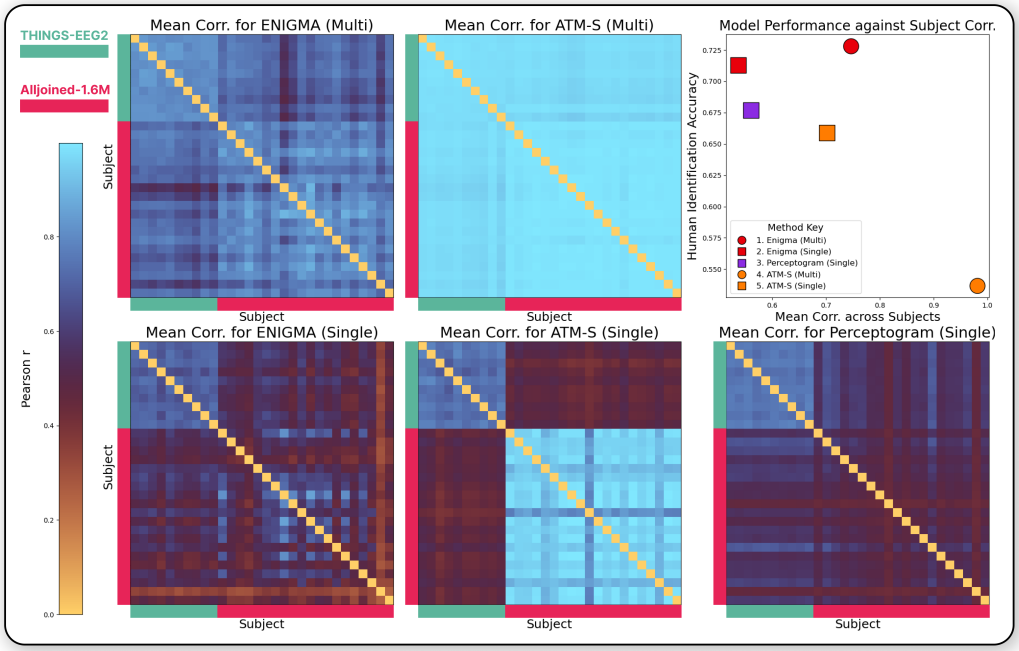

Figure 5: Subject-to-subject Pearson correlation matrices of predicted CLIP embeddings from different subjects in THINGS-EEG2 (green) and Alljoined-1.6M (red). Plots are displayed for all methods in Table 1 in both single and multi-subject configurations. The top right figure plots the human identification accuracy from the behavioral experiment in Section 4.2 (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.

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.

4.4 Additional Analyses

Scaling. 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.

Subject-wise Correlation 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 5. 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. Additional mean correlation plots comparing the x_{EEG} embeddings produced by the subject-specific modules of **ENIGMA** and ATM-S are in Appendix A.6.

5 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 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 desiderate, and lay 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** closes the gap between EEG and fMRI reconstructions by achieving semantically meaningful outputs even from consumer-grade EEG signals on the Alljoined-1.6M dataset. This shift marks a significant step toward deployable 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 a interface for applications of brain decoding technology.

5.1 Current Limitations

Despite training on 30 participants, **ENIGMA's** cross-subject transfer does not introduce any measurable data efficiency benefits, i.e., pre-training does not meaningfully reduce the tens of thousands of trials still required to fine-tune on a new subject. Our investigation into the quality of the CLIP embedding space also raised questions on developing metrics that better respects the non-linear geometry of the manifold. Our model has so far been validated only in a tightly constrained image-reconstruction paradigm, leaving its utility for more open-ended BCI tasks untested. We plan to explore the above research avenues in future work.

5.2 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 [26].

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 affordable brain-computer interfaces. In Review, see Appendix B., 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] 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. URL https://openreview.net/forum?id=paqqd1OODl.

[7] 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).

[8] 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].

[9] 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.

[10] 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.

[11] 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.

[12] 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*.

[13] 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.

[14] 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.

[15] 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.

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

[17] 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].

[18] 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.

[19] 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].

[20] 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.

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

[22] 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.

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

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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 [**?**]. The Emotiv firmware already applies a dual 50/60 Hz notch, effectively attenuating frequencies above 43 Hz, so we added only a 0.5 Hz high-pass and an extra 60 Hz notch to suppress residual line noise. Continuous recordings were then epoched 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 [**? ?**]. Epochs were baseline-corrected to the pre-stimulus window and resampled to 250 Hz to match the ATM-S benchmark [**?**]. Finally, we performed multivariate noise normalization [**?**]: estimating the whitening matrix solely on the training partition and whitening the data improving signal-to-noise ratio from the all train samples on train and test samples to avoid data contamination [2]. This yielded samples $x \in \mathbb{R}^{N_C \times N_T}$ comprising $N_C = 64$ channels for THINGS-EEG2, $N_C = 32$ channels for AJ-1.6, 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 AJ-1.6. For an analysis of this step on performance, see Appendix A.4.

A.2 Additional evaluation metric details

PixCorr is the pixel-level correlation score. SSIM is the structural similarity index metric [27]. AlexNet(2) and AlexNet(5) are the 2-way comparisons (2WC) of layers 2 and 5 of AlexNet [28]. CLIP is the 2WC of the output layer of the CLIP ViT-L/14 Vision model [8]. Incep is the 2WC of the last pooling layer of InceptionV3 [29]. EffNet-B and SwAV are distance metrics gathered from EfficientNet-B13 [30] and SwAV-ResNet50 [31] models. For the metrics in Table 1 of the manuscript, a two-way comparison 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. Two-way identification refers to percent correct across a set of two-way comparisons performed on a pool of distractor images. The two-way identification metrics we report, which are calculated using reconstructions of the 199 other test-set stimuli as distractors, are notably different from the two-way identification 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 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 the implementation of the metrics calculated.

A.3 Statistical significance of metrics

A.4 Channel count ablations

A.5 Behavioral Experiments

A.5.1 Experiment protocols

We conducted a set of behavioral experiments on 545 human raters online. For our experiment, we identified no risks to the human participants, and collected informed consent from all participants. We

probed 1 experiment intermixed with trials from all methods and cases we aimed to evaluate in Table 1, with the experiment consisting of trials sampled evenly from the different stimulus types and the 30 subjects across THINGS-EEG2 and Alljoined-1.6M. The experimental trials were shuffled and 60 trials were presented to each subject. 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. Each human rater was presented with 8 attention check trials during the experiment. An attention check is a trial in which the ground truth image is presented as a candidate image during the trial. Because the ground truth image will always be the image that is most similar to itself, these trials were used to identify whether subjects were paying attention to the task and the instructions. We identified 8 human raters who failed at least 2 attention checks and removed those raters from our data before conducting our analysis. Code to reproduce our experiment can be found in our anonymized GitHub repository.

A.5.2 2AFC identification task

Figure 6: 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 6. 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 scan 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.

A.6 Additional Mean Correlation Plots

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 5.1.

   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 open source code to reproduce our model and our behavioral experiment. A link to an anonymized repository 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 our training hyperparameters can be found in Section 3.3.

   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 A.3.

   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: [Yes]

   Justification: Yes, we discuss the computing resources used in Section 3.3.

   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 Sections 5 and 5.2.

    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 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 model 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 A.5 discusses the protocols of our behavioral experiment and details about compensation.

    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 behavioral experiment provided informed consent before participating, and no risks were posed.

    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: [NA]

    Justification: No LLMs were used in our research.

    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.