ENIGMA: EEG-to-Image in 15 Minutes Using Less Than 1% of the Parameters

Anonymous authors
Paper under double-blind review

Abstract

To be practical for real-life applications, models for brain-computer interfaces must be easily and quickly deployable on new subjects, effective on affordable scanning hardware, and small enough to run locally on accessible computing resources. 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, while fine-tuning effectively on new subjects with as little as 15 minutes of data. **ENIGMA** boasts a simpler architecture and requires less than 1% of the trainable parameters necessary for previous approaches. Our approach integrates a subject-unified spatio-temporal backbone along with a set of multi-subject latent alignment layers and an MLP projector 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 a significant performance boost across both research-grade and consumer-grade EEG hardware, and a substantial improvement in fine-tuning efficiency and inference cost. 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.

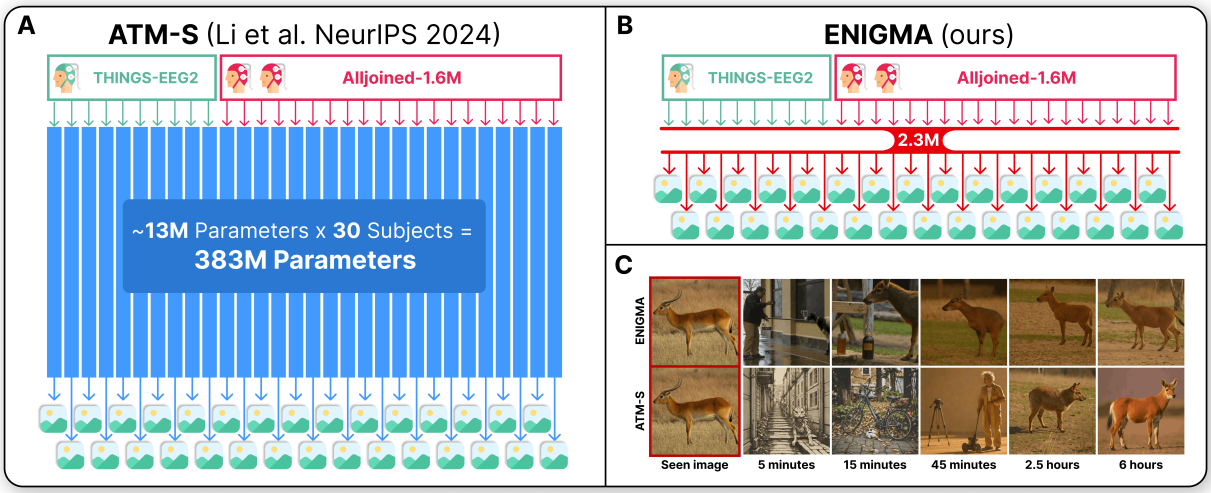

Figure 1: **A**: ATM-S Li et al. (2024) vs **B**: **ENIGMA** (ours) comparison of model size and multi-subject capabilities on THINGS-EEG2 Gifford et al. (2022) (green cap, 10 subjects) and Alljoined-1.6M Xu et al. (2025) (red cap, 20 subjects) datasets. **C**: Comparison of **ENIGMA** and ATM-S across training data scale.

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 targeting states like mental images Kneeland et al. (2025). Decoding visual information encoded in human brain activity could help researchers better understand cognitive processes, and could also be useful in a clinical setting Pearson et al. (2015), where millions of patients are left unable to communicate through conventional means as a result of traumatic brain injuries, and many common afflictions manifest as a profound dysregulation of unwanted or confusing visual experiences Holmes & Mathews (2010). While the Natural Scenes Dataset (NSD) Allen et al. (2022); Kneeland et al. (2025) has yielded striking functional magnetic resonance imaging (fMRI)-based reconstructions of seen images using latent diffusion models Scotti et al. (2024), 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. (2024) demonstrates a simple linear mapping from EEG to an expressive CLIP (Contrastive Language-Image Pretraining Radford et al. (2021)) image embedding space combined with a pre-trained diffusion model, while Li et al. (2024) proposes a more 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. (2023). These approaches have laid important groundwork for this field of research, however, current EEG-to-Image models face three critical barriers that prevent their deployment in practical BCI applications.

Rapid adaptation to new subjects remains an unsolved challenge. All existing methods for EEG-to-Image decoding Song et al. (2024); Fei et al. (2024); Li et al. (2024) require training specialized models from scratch for each subject using hours of training data, which is impractical for real-world BCI applications that demand rapid functionality on new subjects. While recent work in fMRI-to-Image has made progress on this front—Scotti et al. (2024) reduced calibration requirements from 40 hours to 1 hour through efficient fine-tuning—even this breakthrough remains insufficient for practical deployment, as it still requires both an expensive 7 tesla fMRI scanner and a full hour of subject-specific data collection. Critically, no EEG-to-Image model has attempted to enable rapid fine-tuning on new subjects, despite EEG's advantages in accessibility and portability. We hypothesize that this gap has persisted because of EEG's noisy spatio-temporal signal characteristics, making it difficult to leverage generalizable knowledge learned from other subjects while quickly adapting to the idiosyncratic neural patterns of a new individual.

Current approaches lack robustness when applied to affordable hardware. The recent release of the Alljoined-1.6M dataset—designed for evaluating EEG-to-Image models on consumer-grade hardware—revealed a troubling pattern: many recent models, especially complex architectures such as ATM-S, are not robust to drops in hardware quality and fail at a much higher rate when deployed in noisier recording environments Xu et al. (2025). This brittleness represents a fundamental obstacle to democratizing BCI technology, as the vast majority of potential users and downstream applications cannot access research-grade EEG equipment. The challenge lies in learning representations that capture the underlying neural encoding of visual information in a way that generalizes across varying levels of signal quality, electrode density, and noise characteristics. Without addressing this weakness, EEG-to-Image models will remain confined to laboratory settings, unable to translate to the accessible, consumer-grade hardware that many practical BCI applications require.

Existing architectures are too computationally prohibitive for large-scale deployment. Most EEG-to-Image models are quite large relative to the amount of data they are trained on, and require separate models to be trained for each subject, resulting in a linear increase in what is already a very large functional model size as models are deployed across multiple users Song et al. (2024); Fei et al. (2024). While Li et al. (2024) demonstrated that a unified model can be trained across multiple subjects, doing so without any mechanism for handling subject-specific differences led to a substantial performance drop; attaining reasonable performance still required training separate models for each subject. These architectural and size limitations constitute a stubborn barrier for widespread deployment: running these large models on edge devices or in settings without access to server-grade GPUs is almost impossible, and supporting inference on multiple subjects requires scaling computational resources proportionally. The core difficulty lies in designing

a model architecture with performance *and* efficiency in mind, and in ensuring that a given approach can maintain subject-specific decoding performance while sharing the vast majority of parameters across subjects.

In this paper, we introduce **ENIGMA** (**E**EG **N**eural **I**mage **G**enerator for **M**ulti-subject **A**pplications), a multi-subject model for reconstructing seen images from EEG data that directly addresses each of these three critical weaknesses. Our work includes several notable contributions:

(1) ENIGMA is the first EEG-to-Image model to quickly fine-tune on new subjects with as little as 15 minutes of data, making it effective for practical downstream use cases (Figure 1C). **(2)** Our model achieves robust performance across hardware quality levels, producing SOTA performance on both available benchmark datasets, and a substantial performance gain on Alljoined-1.6M, collected onconsumer-grade hardware. **(3) ENIGMA** is unified across subjects and datasets, requiring less than 1% of the trainable parameters to decode the 30 subjects across THINGS-EEG2 and Alljoined-1.6M datasets (5.5x reduction vs ATM-S $\times$ 30 subjects = 165x reduction in total parameters) when compared to single subject approaches (Figure 1AB). **(4)** We are the first EEG-to-Image work to provide extensive evaluations using behavioral experiments with human raters, a standard in the adjacent field of fMRI-to-Image research, and we conduct a detailed ablation analysis investigating which architectural aspects are most effective across differences in hardware quality and multi-subject configurations.

Our work demonstrates significant progress towards overcoming the three fundamental barriers that have prevented EEG-to-Image decoding from practical deployment. By enabling rapid fine-tuning with minimal data, maintaining robust performance across hardware configurations, and achieving dramatic improvements in parameter efficiency, **ENIGMA** represents a sizable step on the path to bring real-time, accessible "mind-to-image" translation from laboratory demonstration to practical clinical, consumer, and research applications.

2 Related Work

fMRI-to-Image Reconstruction. The advent of generative diffusion models has revolutionized neural decoding for adjacent scanning modalities like fMRI Ozcelik & VanRullen (2023); Scotti et al. (2023; 2024); Takagi & Nishimoto (2023a;b); Kneeland et al. (2023b;a;c). Takagi & Nishimoto (2023a) demonstrated some of the first high-resolution image reconstructions from fMRI by mapping fMRI activity into the latent space of a diffusion model. Ozcelik & VanRullen (2023) 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 Ozcelik & VanRullen (2023); Scotti et al. (2023; 2024); Takagi & Nishimoto (2023a;b); Kneeland et al. (2023b;a;c). 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.

Multi-Subject fMRI Models. Recent work in fMRI-to-Image decoding has demonstrated the effectiveness of multi-subject pretraining for rapid adaptation to new subjects. Scotti et al. (2024) introduced MindEye2, which employs a shared-subject functional alignment approach where subject-specific ridge regression linearly maps each individual's fMRI voxel patterns to a common 4096-dimensional latent space, followed by a shared non-linear pipeline across all subjects. By pretraining on 7 subjects and fine-tuning on new subjects, MindEye2 achieves high-quality reconstructions with as little as 1 hour of calibration data—a 40x reduction compared to single-subject approaches. However, such linear alignment strategies on the input data do not translate well to EEG-to-Image decoding due to fundamental differences in signal characteristics: EEG exhibits complex spatio-temporal dynamics with highly variable signal topology across subjects, lower spatial resolution, and substantially higher noise levels compared to the relatively stable spatial patterns of fMRI voxel activity. These challenges necessitate different architectural approaches for achieving robust multi-subject EEG decoding, which we address in this work.

EEG-based Visual Decoding. Decoding visual content from EEG recordings uses a wide array of approaches spanning classification, retrieval, and image reconstruction. While early studies collected EEG recordings in response to visual stimuli Spampinato et al. (2017), many contained confounds that made it difficult to decode true semantic content Li et al. (2020), highlighting a need for better data and methods. This critique, along with improvements in EEG preprocessing and experimental design, led to the development

of new datasets as part of the THINGS initiative. Grootswagers et al. (2022) released THINGS-EEG (50 subjects, 1,854 concepts) using rapid serial visual presentation, and Gifford et al. (2022) 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 Song et al. (2024); Li et al. (2024) employing contrastive learning between image and EEG features have achieved significant gains in zero-shot object classification. Alljoined-1.6M Xu et al. (2025) 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.

EEG-to-Image Reconstruction. In the space of EEG-to-Image reconstruction methods, Li et al. (2024) 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 model. The complexity of the ATM-S architecture requires careful tuning of many intricate architectural components and multiple sequential training stages to be successful. 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 Xu et al. (2025). 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 & VanRullen (2023), Perceptogram (Fei et al., 2024) 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 its reconstructions still contain less detail than fMRI-based reconstruction, the approach of (Fei et al., 2024) demonstrates the significant power of robust linear models in producing recognizable images from low SNR brain activity patterns.

3 ENIGMA

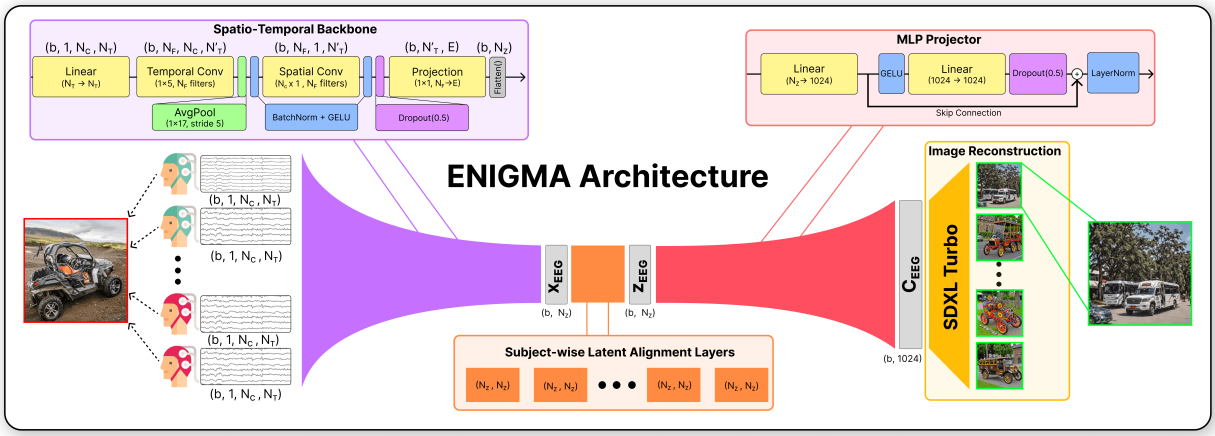

Figure 2: During training, brain activity from each subject is passed through a shared pathway of spatio-temporal convolutions, producing an intermediate latent vector x_{EEG} from all subjects, this is then passed through a set of subject-specific latent alignment layers to produce an aligned latent embedding z_{EEG}. This latent is passed through a fully connected MLP projection layer to produce the output c_{EEG} vector, which is reconstructed into an image using SDXL. Details of these procedures are provided in Section 3.3.

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 subject-unified model architecture that rapidly fine-tunes on new subjects, and **(3)** a small, scalable, and efficient design that minimizes model complexity and compute requirements for both training and inference.

To meet these requirements, our proposed model, **ENIGMA**, has 4 components: **(1)** a spatio-temporal convolutional neural network to learn a set of robust features from the spatial and temporal dimensions of the input signal, **(2)** a set of subject-wise latent alignment layers to capture and navigate subject-specific differences in the latent space of the model, **(3)** an MLP projector to the ViT-H/14 CLIP Radford et al. (2021) embedding space, and **(4)** a pretrained diffusion model (i.e. SDXL Turbo) that can invert a CLIP embedding into an image.

Unlike previous approaches, our model can operate in three distinct modes: single-subject (trained on one subject at a time), multi-subject (trained on multiple subjects in parallel using shared model weights), and fine-tuned (pretrained on one or more subjects, and fine-tuned on a target subject). In practice **ENIGMA** has near identical performance in the single-subject and fine-tuned conditions when using all available training data (see Figure 4)B, and so we primarily report performance metrics in the single and multi-subject configurations, while examining the fine-tuning efficiency benefits provided by our pretraining approach.

3.2 Datasets

THINGS-EEG2 (Gifford et al., 2022) is the current public benchmark for EEG-based visual decoding. It contains 64-channel ActiChamp (costing ∼$60,000) recordings recorded at 1000Hz 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 Xu et al. (2025) 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 Appendix A.1. When reproducing other methods, we follow their preparation and preprocessing steps.

3.3 Architecture

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

Spatio-Temporal Backbone The preprocessed EEG data is passed through an embedding module that applies convolutions over the temporal and spatial dimensions of our data. We treat multichannel EEG with N_C channels and N_T time points 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:

$$
\begin{aligned}
z_1 &= \text{Conv2D}^{N_F}_{1\times 5}(EEG) \\
z_2 &= \text{AvgPool}_{1\times 17,5}(z_1) \\
h_1 &= \text{GELU}(\text{BN}(z_2))
\end{aligned}
$$

where $h_1 \in \mathbb{R}^{N_F \times N_C \times N'_T}$ with $N'_T = \left\lfloor \frac{N_T - 21}{5} \right\rfloor + 1$ (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:

$$
\begin{aligned}
z_3 &= \text{Conv2D}^{N_F}_{N_C\times 1}(h_1) \\
h_2 &= \text{GELU}(\text{BN}(z_3))
\end{aligned}
$$

where $h_2 \in \mathbb{R}^{N_F \times 1 \times N'_T}$. We apply dropout Srivastava et al. (2014) ($p = 0.5$) for regularization and project the features to 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 obtain a latent embedding vector $X_{EEG} \in \mathbb{R}^{N_z}$ of dimension $N_z = 184$.

Subject-wise Latent Alignment Layers To account for systematic differences between subjects, we learn a set of subject-specific fully-connected linear alignment layers with weights $W_s \in \mathbb{R}^{N_z \times N_z}$ that output an aligned latent representation z_{EEG} across subjects. This alignment module allows downstream modules to

learn a unified mapping to the CLIP embedding space across subjects, and is an important piece in allowing the vast majority of model parameters to be shared across multiple subjects with diverse signal dynamics.

MLP Projector After obtaining the aligned z_{EEG} latent embedding, 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 an EEG-predicted CLIP 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) Sauer et al. (2024), and its associated CLIP ViT-H/14 image-prompt adapter (IP-Adapter) Ye et al. (2023). 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:

$$x_T \sim \mathcal{N}(0, I), \tag{1}$$

$$x_{t-1} = f_\theta(x_t, c_{\text{text}}, c_{\text{EEG}}, t) + \text{noise} \tag{2}$$

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 Functions and Training Following Li et al. (2024), 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 van den Oord et al. (2019); Radford et al. (2021). 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 $\lambda = 0.5$, our overall loss function can be seen in Equation 3, and the full mathematical formulations of our loss functions can be found in Appendix A.2.

$$\mathcal{L} = \text{MSE}(c_{EEG}, f_{CLIP}(\text{image})) \quad + \lambda \cdot \text{InfoNCE}(c_{EEG}, \text{norm}(f_{CLIP}(\text{image}))) \tag{3}$$

ENIGMA was trained in FP32 on an RTX 3090 GPU using the AdamW optimizer with learning rate 3e-4 and batch size 512. In the multi-subject configuration, we train for 150 epochs on the training split of both THINGS-EEG2 and Alljoined-1.6M simultaneously (30 subjects, ∼2M data points averaged for each image to ∼500k training trials). In this configuration, the model takes 5.5 hours to train across all 30 subjects, and 10 minutes to train on a single subject. We note that in practice our model could also be trained on GPUs with as little as 8GB of VRAM and fine-tuned in as little as 2 minutes when using a 15 minute calibration period. Such minimal computational requirements allow **ENIGMA** to be rapidly deployable on edge computing devices, facilitating a wide array of downstream applications.

4 Results

We report results on two of the most recent and prominent EEG-to-Image benchmarks: THINGS-EEG2 Gifford et al. (2022) and Alljoined-1.6M Xu et al. (2025), and evaluate **ENIGMA** against the only available EEG-to-Image baselines for those datasets, Perceptogram Fei et al. (2024) and ATM-S Li et al. (2024). Figure 3 presents a set of the best reconstructed images, comparing our method with available baselines on both available datasets. These examples illustrate typical outcomes: our reconstructions generally capture the correct high-level object, e.g., oranges, sheep, furniture, etc. Perceptograms 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 precise object being decoded. Examples of median and worst-case reconstructions can also be seen in Appendix A.5.

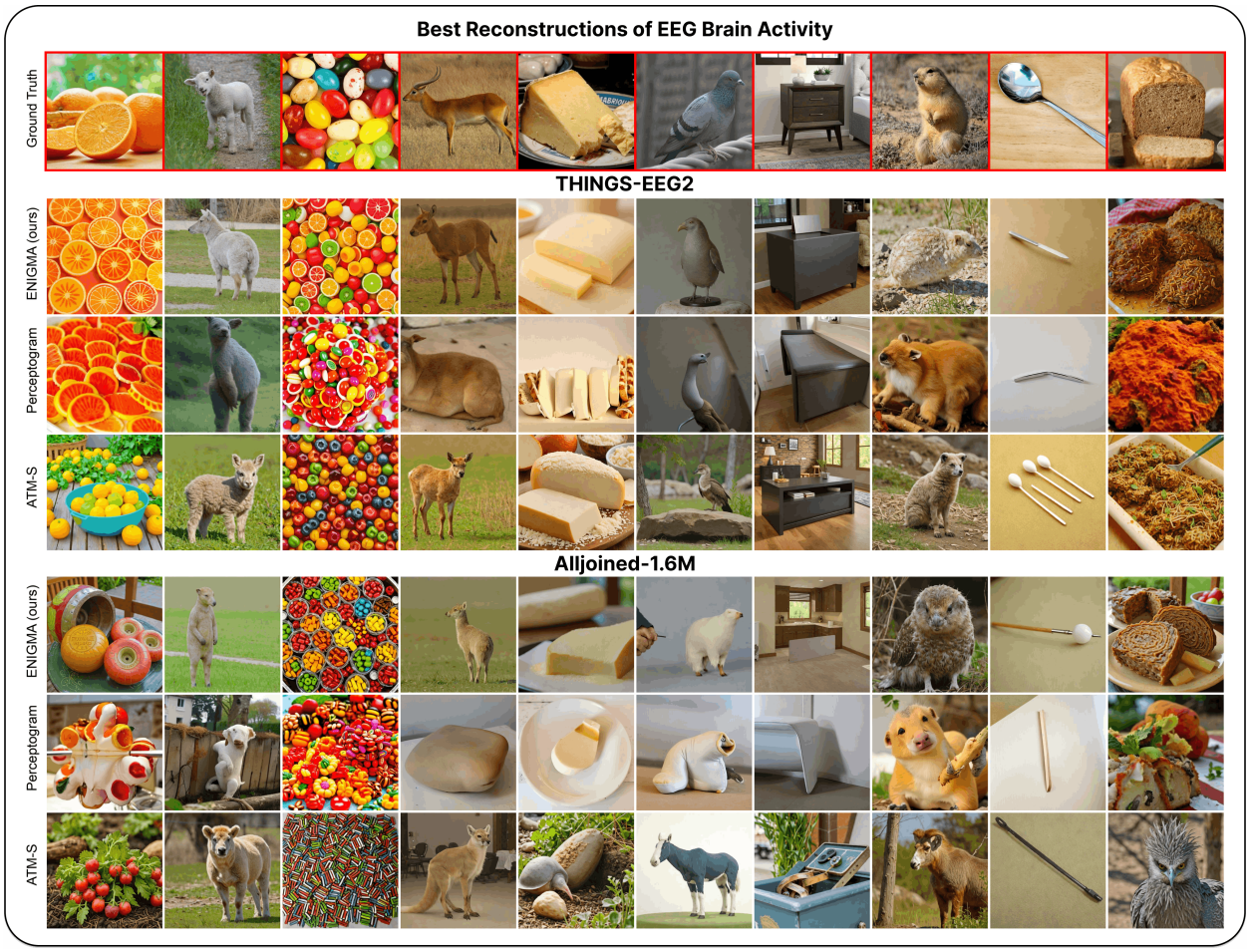

Figure 3: Qualitative comparison of reconstruction methods on seen stimuli from THINGS-EEG2 and Alljoined-1.6M. Reconstructions selected are the outputs sampled from each method and stimulus with the highest scores on all of the image feature metrics in Table 1.

Method	Model Properties		Low-Level		High-Level						Retrieval			Human Raters
	# of Parameters ↓	Inference GFLOPS ↓	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓	Top-1 ↑	Top-5 ↑	Top-10 ↑	Ident. Acc. ↑
THINGS-EEG2 (10 subjects)														
ENIGMA (Multi-Subject)	**2,376,842**	**294.4**	**0.1668**	**0.4264**	**82.99%**	**89.12%**	**76.54%**	**80.33%**	**0.8577**	**0.5399**	**22.55%**	**50.75%**	**64.05%**	**86.04%**
ATM-S (Multi-Subject)	12,815,311	3,858.6	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)	**13,896,820**	**294.4**	0.1718	0.4233	83.64%	**89.49%**	**77.65%**	**81.48%**	**0.8547**	**0.5403**	27.60%	59.35%	71.15%	**86.82%**
ATM-S (Single-Subject)	128,153,110	3,858.6	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	2,807.8	**0.247**	**0.431**	**85.46%**	88.03%	70.40%	71.98%	0.902	0.581	–	–	–	79.17%
Alljoined-1.6M (20 subjects)														
ENIGMA (Multi-Subject)	**2,376,842**	**588.8**	**0.0852**	**0.4175**	**68.33%**	**73.40%**	**63.14%**	**66.38%**	**0.9259**	**0.6127**	**6.00%**	**18.85%**	**28.80%**	**70.74%**
ATM-S (Multi-Subject)	12,765,711	7,717.2	0.068	0.417	53.49%	53.36%	50.72%	51.46%	0.965	0.668	0.73%	4.13%	7.55%	52.18%
ENIGMA (Single-Subject)	**27,793,640**	**588.8**	0.0852	**0.4175**	**68.33%**	**73.40%**	**63.14%**	**66.38%**	**0.9259**	**0.6127**	**6.00%**	**16.25%**	**25.35%**	**71.82%**
ATM-S (Single-Subject)	255,314,220	7,717.2	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	5,615.6	**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 and GFLOPS are computed by adding up the number of parameters and compute necessary to decode all subjects in each dataset (10 subjects for THINGS-EEG2, 20 for Alljoined-1.6M) in each configuration (single vs multi-subject). Details on the human identification accuracy metric are provided in Section 4.2 and Appendix A.8 For the # of parameters, GFLOPS, 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.3.

4.1 Quantitative Evaluations

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 that the latent alignment layers are enabling our model to align representations across subjects from both datasets. For single-subjects, our model still provides SOTA performance on the majority of metrics.

4.2 Human Behavioral Evaluations

For brain decoding models to be deployed in BCI applications, their outputs must be meaningfully interpretable to users, scientists, and clinicians. While many of the metrics in Table 1 are commonly used as proxies for perceptual quality, prior research has established that these automated metrics often fail to align with human assessments of content Sinha & Russell (2011) or quality Kirstain et al. (2023). Human judgment of reconstruction quality is therefore an essential performance metric for evaluating EEG-to-Image models.

To address this, we conducted a carefully controlled, large-scale online behavioral experiment where human raters (n = 545) assessed the quality of reconstructions through 2-alternative forced choice judgments. In each trial, raters determined whether a reconstruction was more similar to its corresponding ground truth image than to a randomly selected reconstruction from the same method, dataset, and subject. This task evaluates whether reconstructions contain meaningful stimulus-specific content—a minimum requirement for practical utility. Detailed experimental protocols are provided in Appendix A.8.

Our results in Table 1 confirm that **ENIGMA** achieves a substantial improvement in human identification accuracy across all evaluated conditions, which we believe to be the most reliable indicator of performance on EEG-to-Image tasks. We also confirm that all reconstruction models evaluated here produce above-chance performance in our randomized control trials ($p < 0.001$).

4.3 Fine Tuning Efficiency

One of the primary goals of our work is to enable high-quality reconstructions from brief calibration sessions suitable for clinical, consumer, and research contexts where extended data collection is infeasible. Figure 4A demonstrates this capability by comparing reconstruction performance across **ENIGMA** models trained or fine-tuned on varying amounts of subject-specific data. The vertical dashed line marks a 15-minute calibration threshold—a reasonable constraint for real-world BCI deployment. While both pretrained and non-pretrained models improve consistently as training data increases, multi-subject pretraining provides substantial advantages in the low-data regime. At the 15-minute threshold (approximately 4,000 training samples), pretrained **ENIGMA** surpasses the fully-trained ATM-S architecture, while the non-pretrained baseline fails to produce identifiable reconstructions. Further analysis of the scaling performance on each benchmark dataset and of the number of EEG channels can also be found in Appendices A.6 and A.7.These results provide empirical evidence that **ENIGMA** and its multi-subject latent alignment procedure enable effective transfer learning in EEG-to-Image decoding, addressing a critical gap where no prior work has demonstrated functional performance on a novel subject with such limited calibration data.

4.4 Ablation Study

To rigorously evaluate EEG-to-Image architectures and critically examine why our method succeeds in multi-subject and consumer-grade contexts where ATM-S breaks down, we conducted ablations on key design choices in our method. Colored numeric identifiers refer to the ablation results in Figure 4B. **ENIGMA** is displayed as (1), [2].

ENIGMA Modules. While linear models do not provide any of the parameter efficiency or fine-tuning benefits of our **ENIGMA** architecture, we find using a well regularized linear backbone [3] in single-subject contexts remains an effective way to decode semantic information from brain activity Ozcelik & VanRullen (2023); Fei et al. (2024). We also find that eliminating the latent alignment layer (4), [5] harms performance

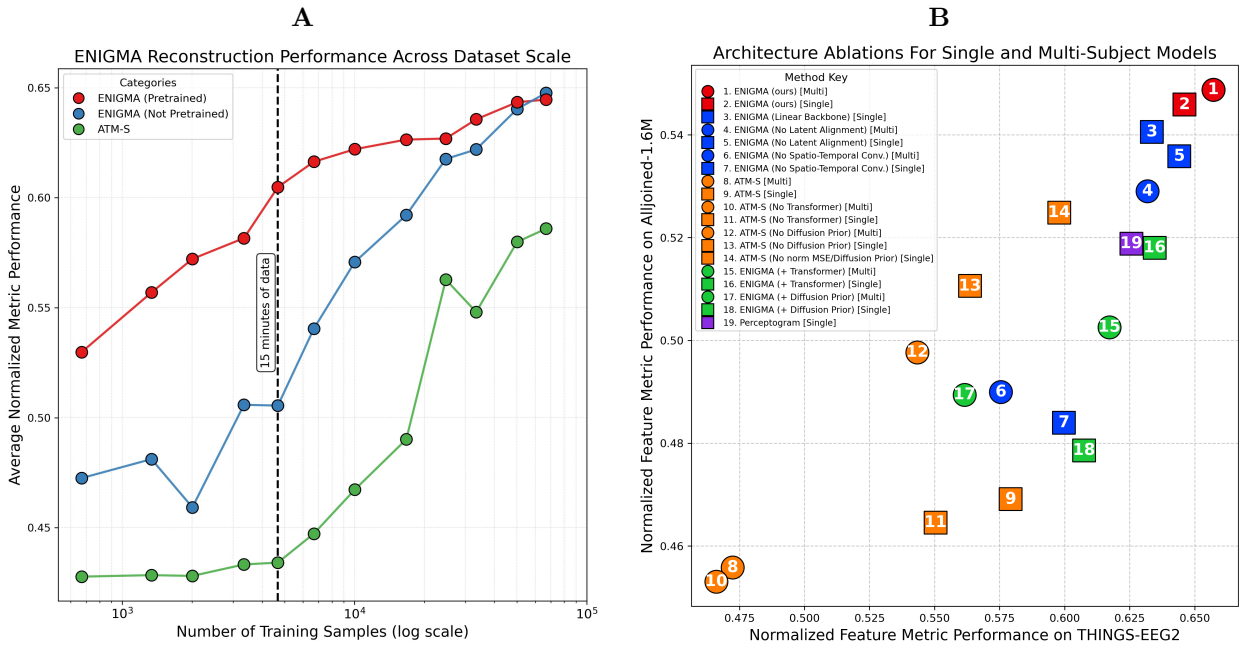

Figure 4: **(A)** Scaling efficiency of **ENIGMA** with (red) and without (blue) pretraining on other subjects, and ATM-S Li et al. (2024) (green). Performance is plotted using varying amounts of target subject training/fine-tuning data on a log-scale X-axis. Reconstruction accuracy, evaluated using the normalized average of feature metrics presented in Table 1, is plotted on the Y-axis. All metrics are calculated on the median subject (2) of the THINGS-EEG2 dataset. **(B)** 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.

disproportionately in multi-subject contexts, showing that the module specifically helps drive 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 EEG brain activity.

ATM-S Modules. While ATM-S's (8), [9] diffusion prior slightly improves its performance on THINGS-EEG2 [13], we find in our analysis that removing it significantly improves performance on data using the consumer-grade EEG hardware in Alljoined-1.6M, and on both datasets in multi-subject contexts (12). We also noticed that if we remove the normalization of the target vectors in the MSE loss computation (see Appendix A.2 for details), results improve even further and negate the need for the diffusion prior entirely [14]. We observe similar patterns with the transformer encoder (10), [11], which only benefits narrow single-subject THINGS-EEG2 settings. These results suggest that these specific architectural complexities hinder both robustness to lower-SNR data and the ability to capture subject-specific variations in unified multi-subject models, informing our simple and robust approach.

ENIGMA + ATM-S Modules. To evaluate whether the architectural modules introduced by Li et al. (2024) would be beneficial to **ENIGMA**, we grafted ATM-S's transformer-based encoder and diffusion prior training stage onto **ENIGMA** (15),[16],(17),[18]. 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 computationally expensive modules.

5 Discussion

Here we introduce **ENIGMA**, a multi-subject EEG-to-Image reconstruction model that directly addresses three critical barriers preventing practical deployment of visual decoding systems: the inability to rapidly fine-tune on new subjects, brittle performance on consumer-grade hardware, and computational prohibitiveness at scale. Through multi-subject latent alignment and a suite of other techniques, **ENIGMA** achieves performant reconstructions with only 15 minutes of calibration data, an order of magnitude reduction compared to existing approaches that require hours of subject-specific training. Our lightweight spatio-temporal CNN architecture maintains robust performance across both research-grade (THINGS-EEG2) and consumer-grade (Alljoined-1.6M) hardware, demonstrating that architectural simplicity and multi-subject pretraining enable effective performance across signals of varying quality. Finally, by sharing model weights across subjects with only lightweight subject-specific latent alignment layers, **ENIGMA** achieves a ∼165x reduction in parameters for multi-subject deployments, enabling efficient simultaneous inference for multiple users with less than 1% of the model size.

Motivated by the overlap between vision and mental imagery Breedlove et al. (2020), we believe models like **ENIGMA** that target semantic content encoded in brain activity from visual perception are a meaningful step to more generalized breakthroughs in "mind reading" technology, and look forward to future research building on our work towards these goals.

5.1 Current Limitations

Despite training on up to 30 participants, **ENIGMA's** multi-subject training and fine-tuning approaches do not introduce any measurable performance gains, i.e., adding more subjects does not meaningfully change the performance ceiling of the model. This aligns with previous findings studying the scaling properties of neuroimaging data Banville et al. (2025), but nonetheless represents an important limitation for the field to overcome. 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 Broader Impacts

While fMRI has dominated neural image reconstruction research due to its superior spatial resolution and downstream decoding performance, its limited accessibility, high cost, and restriction to laboratory settings preclude real-world BCI deployment. **ENIGMA** demonstrates that EEG, despite its lower signal quality, provides sufficient information for semantically meaningful reconstruction when combined with appropriate architectural choices and training strategies. The convergence of rapid adaptation, hardware robustness, and computational efficiency represents a fundamental shift in the practical viability of EEG-based visual decoding, establishing a foundation for deployable brain-computer interfaces in assistive communication, clinical assessment, and real-time neural decoding applications.

5.3 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. Because of these risks, we stress 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 Gordon & Seth (2024).

References

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.

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

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

Jesse L. Breedlove, Ghislain St-Yves, Cheryl A. Olman, and Thomas Naselaris. Generative feedback explains distinct brain activity codes for seen and mental images. *Current Biology*, 30(12):2211–2224.e6, 2020. ISSN 0960-9822. doi: https://doi.org/10.1016/j.cub.2020.04.014. URL https://www.sciencedirect.com/science/article/pii/S0960982220304942.

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.

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).

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.

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.

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

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.

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

Emily A Holmes and Andrew Mathews. Mental imagery in emotion and emotional disorders. *Clinical psychology review*, 30(3):349–362, 2010.

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.

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*, 2023a. doi: 10.32470/CCN.2023.1672-0. URL https://2023.ccneuro.org/view_paper1337.html?PaperNum=1672.

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

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 2023c. URL http://arxiv.org/abs/2306.00927. arXiv:2306.00927 [cs, q-bio].

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.

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 (eds.), *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.

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.

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*.

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.

Joel Pearson, Thomas Naselaris, Emily A Holmes, and Stephen M Kosslyn. Mental imagery: functional mechanisms and clinical applications. *Trends in cognitive sciences*, 19(10):590–602, 2015.

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].

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

Paul S. Scotti, Mihir Tripathy, Cesare Kadir Torrico Villanueva, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A. Norman, and Tanishq Mathew Abraham. Mindeye2: shared-subject models enable fmri-to-image with 1 hour of data. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

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.

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.

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.

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)*, pp. 6809–6818, 2017.

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.

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`.

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*, pp. 14453–14463, 2023a.

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

Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *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*, pp. 6105–6114. PMLR, 2019. URL `http://proceedings.mlr.press/v97/tan19a.html`.

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

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.

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

Jonathan Xu, Ugo Bruzadin Nunes, Wangshu Jiang, Samuel Ryther, Jordan Pringle, Paul S. Scotti, Arnaud Delorme, and Reese Kneeland. Alljoined-1.6m: A million-trial eeg-image dataset for evaluating affordable brain-computer interfaces, 2025. URL `https://arxiv.org/abs/2508.18571`.

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.

A Appendix

A.1 Data Processing and Format

Raw EEG was stored in standard `.edf` files and pre-processed with MNE-Python (Gramfort et al., 2013). The Emotiv firmware applies a dual 50/60 Hz notch by default, effectively attenuating frequencies above 43 Hz, so we added a 0.5 Hz high-pass and an extra 60 Hz notch to suppress residual line noise. We epoched continuous recordings from −200 ms to 1000 ms relative to image onset. Synchronisation glitches in the Emotiv trigger stream led us to discard 0.55–1.12% of trials, which was comparable to exclusion rates reported in earlier Emotiv evaluations (Badcock et al., 2013; Williams et al., 2020). Epochs were then baseline-corrected to the pre-stimulus window and resampled to 250 Hz to match the ATM-S benchmark (Li et al., 2024). Finally, we performed multivariate noise normalization (Guggenmos et al., 2018) where we whiten input data to improve signal-to-noise ratio (SNR). Note that we estimate the whitening matrix only on the training partition to avoid data contamination. This yielded samples $x_{\text{EEG}} \in \mathbb{R}^{N_C \times N_T}$ with $N_C = 64$ channels for THINGS-EEG2, $N_C = 32$ channels for Alljoined-1.6M, and $N_T = 250$ time points for both datasets. For multi-subject models, all models require the same channel count across all subjects, and so for these models we subsample THINGS-EEG2 to the same 32 channels present in Alljoined-1.6M. For an analysis of this step on performance, see Appendix A.7. Both the THINGS-EEG2 and Alljoined-1.6M datasets contain 4 image repetitions per training sample, and 80 repetitions per inference sample. For training and inference, we average together these multiple trials for each image presentation to further boost the SNR of the data.

A.2 Loss function formulations

For the loss functions behind the training of ENIGMA, we use common, well-defined loss functions for mean-squared error and symmetric InfoNCE contrastive loss widely utilized in the literature, with precise mathematical definitions below.

$$L = \text{MSE}(c_{EEG}, f_{CLIP}(\text{image})) + \lambda \cdot \text{InfoNCE}(c_{EEG}, \text{norm}(f_{CLIP}(\text{image}))) \tag{4}$$

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^{N} \| c_{EEG}^{i} - f_{CLIP}(\text{image})^{i} \|_2^2 \tag{5}$$

$$\text{InfoNCE} = \frac{L_{EEG \to Image} + L_{Image \to EEG}}{2} \tag{6}$$

$$L_{EEG \to Image} = -\frac{1}{N} \sum_{i=1}^{N} \log \left[\frac{\exp(c_{EEG}^{i} \cdot f_{CLIP}(\text{image})^{i} / \tau)}{\sum_{j=1}^{N} \exp(c_{EEG}^{i} \cdot f_{CLIP}(\text{image})^{j} / \tau)} \right] \tag{7}$$

$$L_{Image \to EEG} = -\frac{1}{N} \sum_{i=1}^{N} \log \left[\frac{\exp(c_{EEG}^{i} \cdot f_{CLIP}(\text{image})^{i} / \tau)}{\sum_{j=1}^{N} \exp(c_{EEG}^{j} \cdot f_{CLIP}(\text{image})^{i} / \tau)} \right] \tag{8}$$

Here:

- $c_{EEG} \in \mathbb{R}^{1024}$ is the output of our EEG encoding pipeline (after MLP projection)
- $f_{CLIP}(\text{image}) \in \mathbb{R}^{1024}$ is the CLIP image embedding
- $\cdot$ denotes the dot product between normalized embeddings
- $\tau = 1/\text{logit_scale}$ is the learnable temperature parameter
- The diagonal elements ($i = j$) represent positive EEG-image pairs

- All off-diagonal elements serve as negatives within the batch

The symmetric formulation ensures bidirectional alignment between EEG and image modalities. By maximizing $c^i_{EEG} \cdot f_{CLIP}(\text{image})^i$ for paired samples while minimizing similarity with other samples in the batch, we learn a joint embedding space capturing the semantic correspondence between brain activity and visual stimuli.

In our research we also noticed that the MSE loss term of the ATM-S backbone Li et al. (2024) was computed on unit-normalized CLIP targets: $\text{norm}(f_{CLIP}(\text{image}))$, inadvertently optimizing for angular alignment (cosine similarity) rather than Euclidean distance in the embedding space. 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, and note that doing so negates the need for the secondary diffusion prior training stage in Li et al. (2024) that was previously necessary to learn the magnitude of $f_{\text{CLIP}}(\text{image})$ (Figure 4B [14]).

A.3 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 Wang et al. (2004).
- AlexNet(2) and AlexNet(5) are the 2-way comparisons (2WC) of layers 2 and 5 of AlexNet Krizhevsky et al. (2012).
- CLIP is the 2WC of the output layer of the CLIP ViT-L/14 Vision model Radford et al. (2021).
- Incep is the 2WC of the last pooling layer of InceptionV3 Szegedy et al. (2015).
- Eff and SwAV are distance metrics gathered from EfficientNet-B13 Tan & Le (2019) and SwAV-ResNet50 Caron et al. (2020) 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 Allen et al. (2022), 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.4 Statistical Significance of Evaluation Metrics

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

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

A.5 Median and Worst-case Reconstructions

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

Figure 6: Qualitative comparison of reconstruction methods on seen stimuli from Alljoined-1.6M. Reconstructions selected are the outputs sampled from each method and stimulus with the median scores on all of the image feature metrics in Table 1.

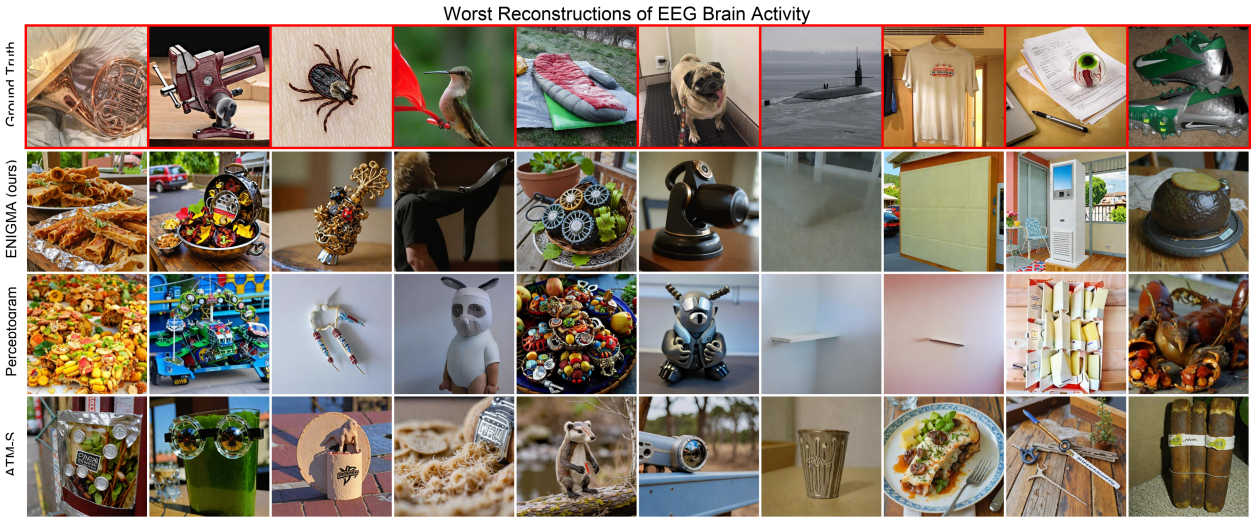

Figure 7: Qualitative comparison of reconstruction methods on seen stimuli from THINGS-EEG2. Reconstructions selected are the outputs sampled from each method and stimulus with the worst scores on all of the image feature metrics in Table 1.

Figure 8: Qualitative comparison of reconstruction methods on seen stimuli from Alljoined-1.6M. Reconstructions selected are the outputs sampled from each method and stimulus with the worst scores on all of the image feature metrics in Table 1.

A.6 Dataset Scaling Analysis

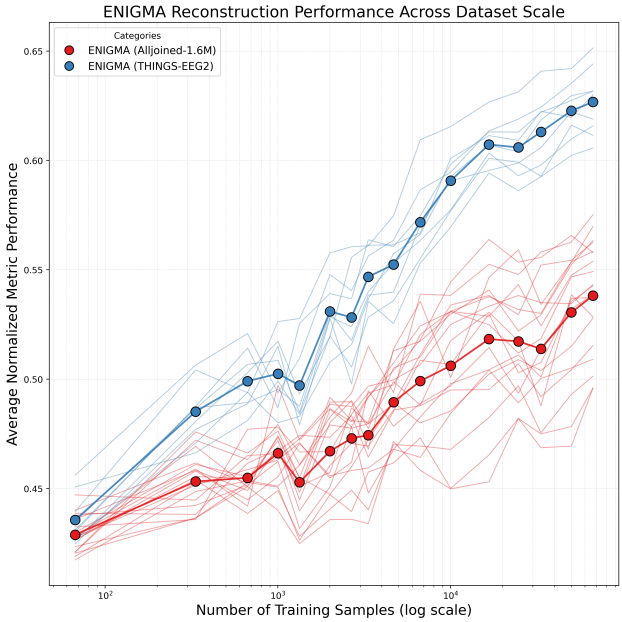

Figure 9: Scaling analysis of **ENIGMA** (not pretrained) performance on the THINGS-EEG2 and Alljoined-1.6M datasets. The number of training samples are plotted on a log-scale X-axis, and the normalized average of feature metrics presented in Table 1 is plotted on the Y-axis for each subject, with the bolded line being the average of all subjects.

When collecting large neuroimaging datasets, it is useful to be able to predict how efficiently models will scale on newly collected data. As shown in Figure 9, the reconstruction performance of **ENIGMA** increases log-linearly with the number of training samples on both available datasets, with no evident saturation on either dataset, however, performance scales with a larger exponent factor on THINGS-EEG2 that was collected on research-grade 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 Xu et al. (2025), this difference in scaling efficiency between EEG hardware quality is a key limitation to overcome for practical BCI applications.

A.7 Channel Count Ablation Analysis

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

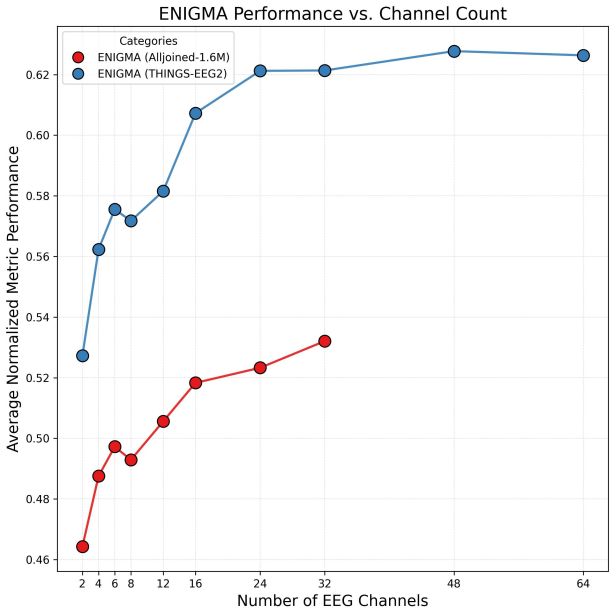

Figure 10: Channel count analysis of model performance for each dataset. The number of channels in each dataset was progressively reduced, while the remaining channels focus primarily on occipital cortex. The Y axis is plotted the same as Fig. 9.

A.8 Behavioral Evaluation Experiments

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

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

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

A.8.1 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.

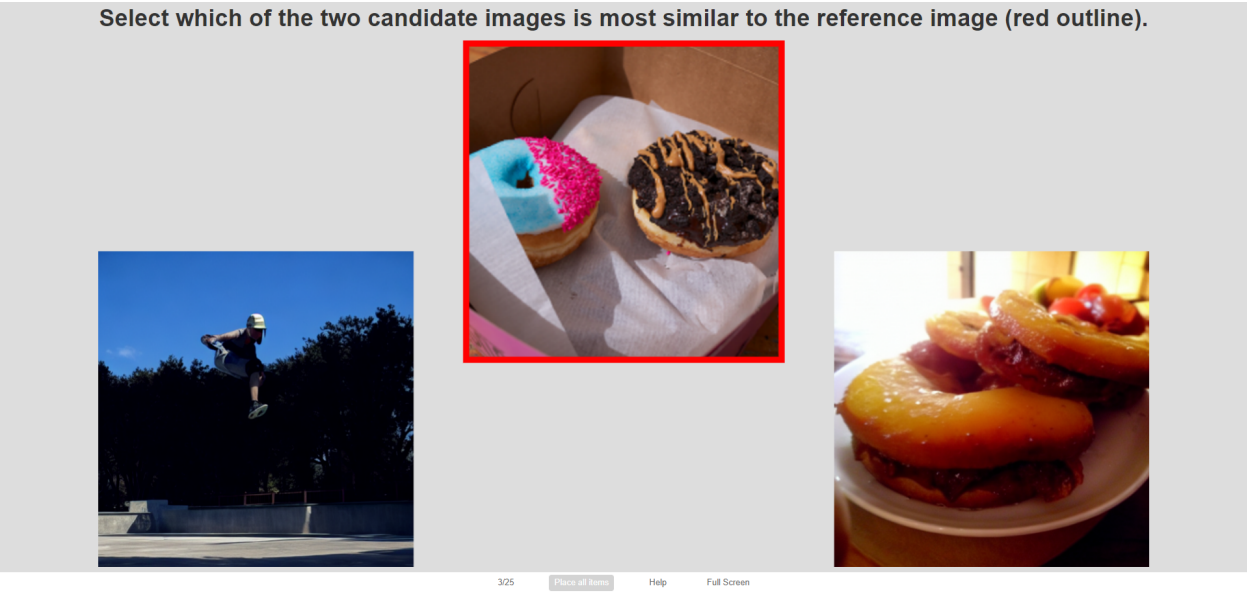

Figure 11: An example of the 2 alternative forced choice task used in our behavioral experiment performed by human raters.