# Brain Image Reconstruction with Retrieval-Augmented Diffusion

Shuqi Zhu
DCST, Tsinghua University
Beijing, China
zsq19991106@gmail.com

Ziyi Ye
DCST, Tsinghua University
Beijing, China
yeziyi1998@gmail.com

Yi Zhong
DCST, Tsinghua University
Beijing, China
zhong-y22@mails.tsinghua.edu.cn

Qingyao Ai*
DCST, Tsinghua University
Beijing, China
aiqy@tsinghua.edu.cn

Yujia Zhou
DCST, Tsinghua University
Beijing, China
zhouyujia@mail.tsinghua.edu.cn

Yiqun Liu
DCST, Tsinghua University
Beijing, China
yiqunliu@tsinghua.edu.cn

## Abstract

Reconstructing visual images from brain signals is a rapidly evolving research with promising applications in brain-computer interfaces, cognitive neuroscience, and assistive technologies. While visual reconstruction based on functional Magnetic Resonance Imaging (fMRI) has previously achieved notable success, this paper explores cost-effective brain signals, i.e., electroencephalography (EEG) and magnetoencephalography (MEG). These signals are less precise than fMRI, which presents greater challenges for reconstruction. To address this problem, we propose **BReAD** (**B**rain Image Reconstruction with **Re**trieval-**A**ugmented **D**iffusion), a novel framework that combines EEG/MEG signals with retrieval-augmented diffusion models to improve image reconstruction quality. BReAD utilizes the semantics decoded from brain signals for (1) retrieving semantic priors from a large-scale image database and (2) serving as a conditional constraint during the diffusion process. Extensive experiments demonstrate that BReAD significantly outperforms existing approaches in both qualitative and quantitative evaluations, paving the way for more robust and practical brain-to-image reconstruction systems. Our codes are available at https://anonymous.4open.science/r/BReAD-anonymous-6457.

## CCS Concepts

• **Human-centered computing** → **Human computer interaction (HCI)**; • **Information systems** → **Information retrieval**; • **Computing methodologies** → *Neural networks*; **Reconstruction**.

## Keywords

BCI, Multimodal, Brain Image Reconstruction, Diffusion Model, Retrieval-Augmented Generation

*Corresponding author.


*SIGIR '25, July 13–18, 2025, Padua, Italy*

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

**ACM Reference Format:**
Shuqi Zhu, Ziyi Ye, Yi Zhong, Qingyao Ai, Yujia Zhou, and Yiqun Liu. 2018. Brain Image Reconstruction with Retrieval-Augmented Diffusion. In *Proceedings of The 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25).* ACM, New York, NY, USA, 11 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 Introduction

Reconstructing human visual perception from brain signals represents a significant frontier in neuroscience, offering potential breakthroughs in brain-computer interface (BCI) systems, cognitive research, and assistive technologies. By decoding neural activity into visual outputs, a series of revolutionary applications have emerged, such as enhancing communication for individuals with locked-in syndrome, and deepening our understanding of how the brain processes visual information [2, 4, 33]. Despite these promises, achieving high-quality reconstructions that are semantically consistent with human perception remains a formidable problem, primarily due to the complexity and variability of brain signals.

To address this problem, generative models, represented by generative adversarial network (GAN) and the diffusion model, have significantly advanced the performance of traditional methods in this task. Such advancement is particularly remarkable with the brain input of functional magnetic resonance imaging (fMRI) signals [11, 20, 25, 26, 31]. fMRI has a higher spatial resolution than other brain signals such as electroencephalography (EEG) and magnetoencephalography (MEG). However, fMRI has significant limitations: including expensive and complex equipment, low temporal resolution, and significant delay, mitigating its usage in a lot of real-time scenarios. On the other hand, while EEG and MEG are more cost-effective and offer high temporal resolution for real-time applications, using these signals to reconstruct images is still at a preliminary stage [34]. Previous attempts to transfer the same schema of generative models to EEG signals [2, 16, 18, 28, 39] often failed to reach a feasible performance in comparison to those reconstructed from fMRI data. In contrast, another line of studies tries to build retrieval models with EEG signals to retrieve semantically relevant images [16, 27**?** ], which is more feasible than reconstruction with only EEG signals. However, such a method restricts the "reconstruction" results to a fixed retrieval dataset, which may not accurately reflect the human's visual perception.

To tackle the above challenges, we propose **BReAD** (**B**rain Image Reconstruction with **Re**trieval-**A**ugmented **D**iffusion), a framework

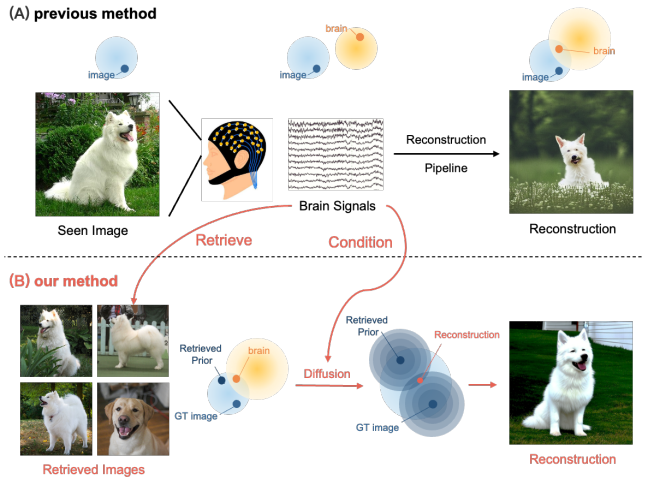


**Figure 1: Illustration of the proposed method compared to the previous method. While previous methods directly map brain signals to the latent space and decode the result, our framework retrieves semantically relevant priors and refines them through a conditioned diffusion process guided by brain embeddings before decoding into final reconstruction.**

for visual reconstruction from EEG/MEG signals with a specially designed retrieval-augmented diffusion model. Figure 1 shows how our method differs from previous approaches by introducing a retrieval-augmented diffusion process. The design of BReAD is motivated by the fact that the human brain leverages prior knowledge and contextual information for forming ideas and hypotheses. BReAD employs a three-stage process to generate images based on EEG signals: (1) Brain signals collected by EEG/MEG devices are encoded into an embedding space shared with a pre-trained image encoder, enabling shared and aligned representation between neural modalities and visual modalities. (2) Relevant images are retrieved based on the encoded brain signal representations from a large-scale image dataset using a similarity-based search, providing contextual priors for the reconstruction. (3) Inspired by SDEdit [17], we refine the retrieved priors through a diffusion pipeline, where a forward diffusion process introduces stochastic noise and a reverse diffusion process iteratively generates high-quality images conditioned on both the priors and brain embeddings. Experimental results show that BReAD significantly outperforms existing methods across a wide range of metrics including pixel-level accuracy, structural similarity, and semantic alignment. These findings validate the effectiveness of our approach in producing images that are both visually detailed and semantically consistent. The primary contributions of this work are as follows:

- We propose BReAD, a novel framework that reconstructs images from EEG/MEG signals with a retrieval augmented diffusion method. Our approach significantly improves the quality of reconstructed images by leveraging brain input for both the semantically meaningful priors and the conditional generation of diffusion processes.
- We extensively evaluated our framework on EEG-ImageNet, ThingsEEG, and ThingsMEG datasets, demonstrating its effectiveness across multiple brain signal modalities.
- We present in-depth analyses showing that the retrieval-augmented approach effectively expands the capabilities of models when the retrieval dataset is incomplete or insufficient to represent human semantics. Our analysis also revealed a positive correlation between retrieval performance and reconstruction quality, showcasing the importance of leveraging retrieval priors for improved semantic alignment and visual fidelity.

## 2 Related Work

### 2.1 Brain Image Reconstruction

Recent advancements in brain image reconstruction have explored various methodologies for mapping neural activity to visual representations. In the context of fMRI, Gu et al. [11] proposed a surface-based convolutional network to decode natural image stimuli, highlighting the role of cortical surface representations in improving reconstruction fidelity. Takagi and Nishimoto [31] introduced a latent diffusion model framework to generate high-resolution images from fMRI signals, demonstrating its ability to capture intricate visual details. Similarly, Ozcelik and VanRullen [20] utilized a generative latent diffusion approach to reconstruct natural scenes from fMRI data, focusing on semantic coherence in the generated outputs. Scotti et al. [25] proposed Mind's Eye, a framework combining contrastive learning with diffusion priors to enhance fMRI-to-image reconstructions, achieving a balance between low-level visual details and high-level semantic alignment. Scotti et al. [26] also introduced a shared-subject modeling approach in MindEye2, enabling effective reconstruction with minimal data by mapping brain activity to a shared latent space. Additionally, Xie et al. [36] developed BrainRAM, a retrieval-augmented framework that integrates priors from large-scale databases at embedding-level to improve semantic consistency and visual quality. The key difference between BrainRAM and our work is that in our approach, the retrieved images are directly used in the diffusion model's generation process, rather than being implicitly involved at the embedding level. By initializing the diffusion process with retrieved priors, our method allows for more accurate and detailed image reconstructions.

For EEG and MEG-based reconstruction, Singh et al. [28] proposed EEG2IMAGE, a framework for synthesizing images from EEG signals, showcasing its feasibility for non-invasive brain signal decoding. Mishra et al. [18] introduced NeuroGAN, an attention-based GAN architecture that effectively captures spatial and temporal features in EEG signals. Zeng et al. [39] developed DM-RE2I, a diffusion model-based framework for reconstructing high-quality images from EEG data. Benchetrit et al. [3] focused on real-time reconstruction of visual perception, emphasizing the potential for real-time decoding applications. Li et al. [16] used guided diffusion processes to enhance alignment between EEG embeddings and reconstructed images, achieving improved semantic fidelity. Bai et al. [2] introduced DreamDiffusion, leveraging temporal masked signal modeling and CLIP alignment to produce high-quality EEG-to-image reconstructions, effectively capturing both temporal dynamics and semantic content. In contrast, our work innovatively

introduces retrieval-augmented diffusion into the reconstruction process, where retrieved images act as priors that directly participate in guiding the diffusion model's generation process. This integration allows the model to leverage both the semantic information from neural signals and the detailed structure provided by the retrieved images. Additionally, our framework is supported by extensive experiments, which include both qualitative and quantitative evaluations. These provide a solid foundation for future comparisons and further advancements in the field.

## 2.2 Diffusion Model

Diffusion models [8] has become a cornerstone in generative modeling, particularly for image synthesis and editing tasks. These models operate by progressively adding noise to data and then learning to reverse this process, effectively modeling complex data distributions. Latent Diffusion Models (LDMs) [23] perform the diffusion process in a compressed latent space rather than directly on high-dimensional pixel data, substantially reducing computational requirements while maintaining high-quality outputs. Meng et al. [17] introduced SDEdit, which employs stochastic differential equations to refine noisy inputs into realistic outputs, making it particularly effective for tasks such as sketch-based image editing. Additionally, Blattmann et al. [5] proposed Retrieval-Augmented Diffusion Models (RADMs), which enhance generation quality by retrieving semantically relevant priors from an external database and conditioning the diffusion process on these priors. Inspired by these approaches, we use retrieved images as priors in the diffusion model to construct the BReAD framework, aiming to improve visual reconstruction quality from neural signals. This integration enables the model to leverage the semantic richness of retrieved priors and the generative power of diffusion processes to achieve high-quality image reconstructions.

# 3 Method

In this section, we first formalize the task of brain image reconstruction. Then we present the methodology of our proposed BReAD framework for reconstructing visual images from brain signals, as illustrated in Figure 2. The method consists of three key components: ***Brain Encoder***, ***Brain Image Retrieval***, and ***Diffusion Pipeline***. First, the Brain Encoder transforms raw brain signals into embeddings aligned with the image embedding space, enabling a shared representation for both modalities. Second, the Brain Image Retrieval module retrieves semantically similar images from a large-scale image dataset as priors for the generation process. Finally, the Diffusion Pipeline refines these priors with semantic information extracted from the brain embeddings into high-quality reconstructed images through an iterative denoising process. Together, these components establish a robust framework for brain-based image reconstruction.

## 3.1 Problem Formulation

The problem investigated in our study aims to reconstruct visual stimuli from human participants' corresponding brain signals. These stimuli are carefully selected to represent a variety of visual categories (e.g., animals, objects, scenes) and are presented to the participant during the data collection phase. Visual stimuli are typically displayed on a screen while the participant's brain activity is measured using electrodes placed on the scalp. Brain signals record the brain's electrical/magnetic activity collected by EEG/MEG devices in response to these stimuli. These signals, often in the form of multi-variate time series, are processed to extract meaningful features, which are then used to predict corresponding visual stimuli.

Consider the dataset $\Omega = \{S_i, V_i\}_{i=1}^{N}$, where $S_i \in \mathbb{R}^{N_c \times D}$ represents the preprocessed brain signals, and $V_i \in \mathbb{R}^{H \times W \times 3}$ denotes the corresponding visual stimuli image. Here, $N_c$ is the number of channels in the brain signal, $D$ is the feature dimension for each channel obtained through preprocessing. $H$, and $W$ are the height and width of the 3-channel RGB image, respectively. Our goal is to map the brain's electrical activity, captured as brain signals $S_i$, back to its corresponding visual stimulus image $V_i$.

## 3.2 Brain Encoder

The brain encoder $\mathcal{F}$ is trained to project the brain signal $S_i$ into the shared embedding space of stimuli images, enabling alignment with the image embeddings derived from $V_i$. This alignment enables the effective integration of neural signals and visual priors in the subsequent retrieval and generation processes. The brain encoder uses a multi-layer network architecture and is trained using a joint loss function that combines MSE loss and InfoNCE loss to align the embeddings of brain signals and images effectively. Next, we define the brain signal embeddings $z_{S_i} = \mathcal{F}(S_i)$ and the corresponding stimuli (i.e., ground truth) image embeddings $z_{V_i} = \mathcal{E}(V_i)$, where $\mathcal{E}$ denotes the image encoder used to extract semantic image features. The MSE loss ensures that the brain signal embeddings $z_{S_i}$ align closely with the corresponding image embeddings $z_{V_i}$:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^{N} |z_{S_i} - z_{V_i}|_2^2 \tag{1}$$

Since relying solely on MSE loss can result in embedding representations lacking discriminative ability [25] , we add a contrastive learning target through the InfoNCE loss[19]. For a given embedding pair $(z_{S_i}, z_{V_i})$, the InfoNCE loss is formulated as:

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(z_{S_i} \cdot z_{V_i}/\tau)}{\sum_{j \neq i} \exp(z_{S_i} \cdot z_{V_j}/\tau)} \tag{2}$$

Here, $\tau$ is a temperature hyperparameter controlling the sharpness of the distribution. For each brain signal embedding $z_{S_i}$, the corresponding image embedding $z_{V_i}$ is treated as the positive example, while the embeddings of other images in the same batch $\{z_{V_j}\}_{j \neq i}$ serve as negative examples, ensuring efficient contrastive learning within the batch. The total loss is defined as a weighted sum of these two components:

$$\mathcal{L} = (1 - \lambda)\mathcal{L}_{\text{MSE}} + \lambda \mathcal{L}_{\text{InfoNCE}} \tag{3}$$

where $\lambda$ is a weighting parameter that balances the contribution of the two terms. By combining MSE loss and InfoNCE loss, the brain encoder not only aligns the brain signal to the corresponding image representations, but also maximizes its separability of non-matching images. This ensures that the learned embeddings capture both intra-class similarity and inter-class distinction, enhancing the

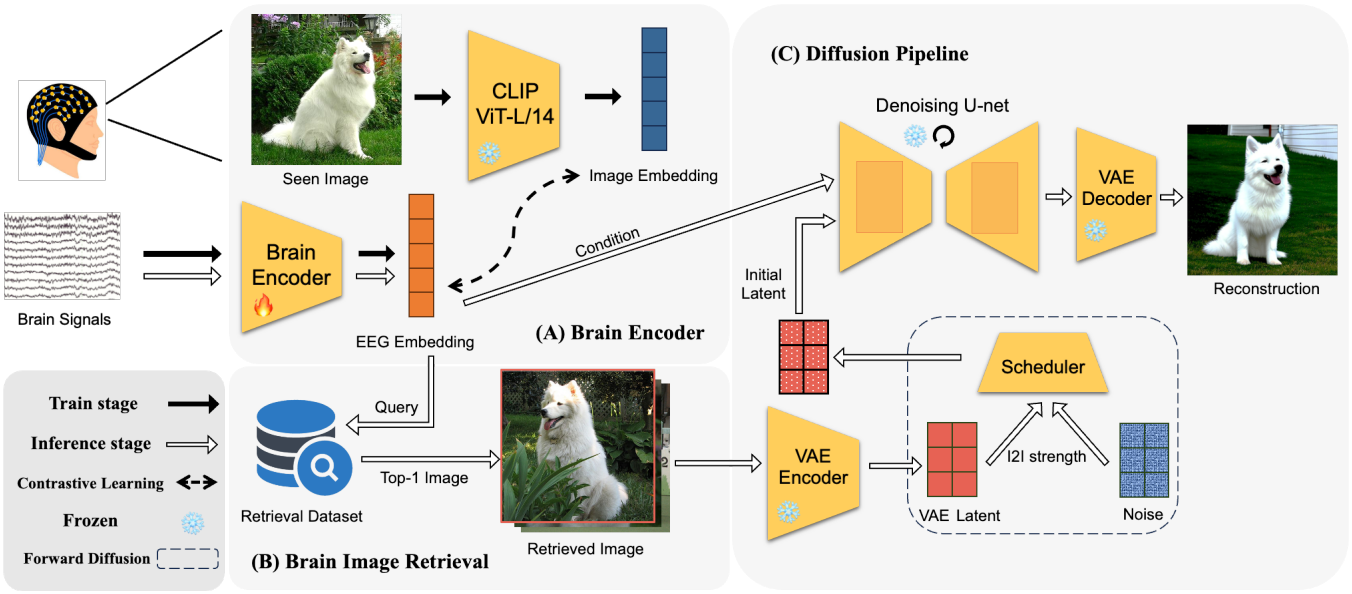


**Figure 2: The main procedure of the proposed framework BreAD. (A) Brain Encoder transforms brain signals into embeddings aligned with a shared representation space using contrastive learning. (B) Brain embeddings are used in the Brain Image Retrieval module to retrieve semantically relevant images from a large-scale image dataset as priors for image reconstruction. (C) The retrieved priors are inputted into a Diffusion Pipeline, where forward diffusion introduces noise to balance prior information with flexibility, and reverse diffusion iteratively generates high-quality reconstructed images conditioned on the brain embeddings and retrieved priors.**

representation alignment and uniformity of the shared embedding space.

## 3.3 Brain Image Retrieval

The brain image retrieval module is designed to enhance the generative process by leveraging semantic priors retrieved from a large-scale image dataset $C = \{C_i\}_{i=1}^{M}$, where $C_i$ denotes the candidate images in the retrieval dataset. The retrieval is achieved by calculating similarity scores between the brain embedding $z_{S_i}$ and the image embeddings from the dataset $C$. We perform the retrieval using Approximate Nearest Neighbor (ANN) search, which is crucial for efficiently retrieving images from a large dataset without performing exhaustive pairwise similarity calculations. The ANN search allows us to quickly identify the nearest neighbors in the embedding space, significantly reducing computational complexity. The similarity between brain signal embeddings and image embeddings is computed using cosine similarity:

$$\text{sim}(z_{S_i}, z_{C_j}) = \frac{z_{S_i} \cdot \mathcal{E}(C_j)}{|z_{S_i}||\mathcal{E}(C_j)|} \tag{4}$$

This results in a set of top-K semantically similar images that serve as candidate priors for the generative process. We select the top-1 image $C_{rel}$ for the subsequent steps.

The retrieved images provide structural and contextual guidance, ensuring that the reconstruction is anchored to meaningful visual features aligned with the neural activity. To prepare these retrieved images for the generative process, they are processed through a pre-trained VAE encoder to generate latent representations:

$$\mathbf{z}_0 = \mathcal{E}_{VAE}(C_{rel}) \tag{5}$$

The diffusion pipeline is then initialized by these latent representations, which significantly reduces the ambiguity and uncertainty inherent in direct generative approaches.

## 3.4 Diffusion Pipeline

The diffusion pipeline serves as the core generative model of the framework, responsible for refining the retrieved priors into high-quality reconstructed images conditioned on the brain signal embedding. The diffusion model we used in this pipeline is based on a latent diffusion model (LDM) [23], which operates in a compressed latent space instead of the pixel space, significantly improving computational efficiency without compromising quality.

The pipeline begins by introducing noise into the retrieved priors through a forward diffusion process. This step is inspired by the SDEdit framework (Stochastic Differential Editing) [17]. The motivation for this forward diffusion is to perturb the retrieved priors by adding controlled amounts of noise, effectively projecting them back into the latent space where the generative model can better explore alternative visual possibilities. By doing so, the generative process avoids over-reliance on the initial priors, ensuring that the final reconstructed image captures both the semantic information from the priors and the unique constraints provided by the brain embedding. The forward diffusion process is mathematically described as follows:

$$q(\mathbf{z}_t|\mathbf{z}_0) = \mathcal{N}(\mathbf{z}_t; \sqrt{\alpha_t}\mathbf{z}_0, (1-\alpha_t)\mathbf{I})$$
$$\mathbf{z}_t = \sqrt{\alpha_t}\mathbf{z}_0 + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}) \tag{6}$$

where $\mathbf{z}_0$ represents the VAE latent of the retrieved image prior, $\alpha_t$ is a noise scheduling parameter, and t is the current diffusion step, controlled by a hyperparameter called img2img strength (I2I strength). $q(\mathbf{z}_t|\mathbf{z}_0)$ represents the conditional probability distribution of $\mathbf{z}_t$ under the condition of $\mathbf{z}_0$. Over a fixed number of steps, this process introduces progressively more noise, resulting in a noisy latent representation $\mathbf{z}_t$ that serves as the starting point for the reverse diffusion process.

The reverse process then iteratively denoises the latent representation from $\mathbf{z}_t$ back to the representation $\hat{\mathbf{z}}_0$, incorporating the brain embedding as conditions. The reverse denoising process is guided by a generative model, such as a U-Net [24], and can be expressed as:

$$p_\theta(\hat{\mathbf{z}}_{t-1}|\hat{\mathbf{z}}_t, \mathbf{z}_{S_i}) = \mathcal{N}(\hat{\mathbf{z}}_{t-1}; \mu_\theta(\hat{\mathbf{z}}_t, t, \mathbf{z}_{S_i}), \Sigma_\theta(\hat{\mathbf{z}}_t, t))$$
$$\hat{\mathbf{z}}_{t-1} = \hat{\mathbf{z}}_t - \eta \cdot \nabla \hat{\mathbf{z}}_t \log p(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_{t+1}) \tag{7}$$

where $\theta$ represents model parameters (guidance scale and so on), $\mu_\theta$ and $\Sigma_\theta$ are the predicted mean and variance calculated by generative model, respectively. $\eta$ denotes the step size of the denoising process. The denoising process is carried out by minimizing the following loss:

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{t,\hat{\mathbf{z}}_0,\boldsymbol{\epsilon}}\left[||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, \mathbf{z}_{S_i})||^2\right] \tag{8}$$

where $\boldsymbol{\epsilon}_\theta$ is the noise prediction function. After the reverse diffusion process completes, the refined latent $\hat{\mathbf{z}}_0$ is decoded back into the image space using the same pre-trained VAE decoder.

The U-Net model is conditioned on the brain signal embeddings, allowing it to incorporate the neural information at every step of the reverse process. Additionally, the retrieved priors provide structural guidance. By performing forward diffusion, the pipeline ensures that the reconstruction process leverages the flexibility and generative capacity of the diffusion model while still being strongly informed by the retrieved priors and brain embeddings. This two-step process—perturbing the priors and then iteratively refining them—enables the generation of high-quality images that are aligned with the semantics in the underlying brain activity signals.

# 4 Experimental Setup

## 4.1 Dataset and Preprocessing

We utilize three datasets for the experiment, including two EEG datasets EEG-ImageNet [41] and Things-EEG2 dataset [10], and a MEG dataset Things-MEG [12]. EEG-ImageNet is collected from 16 participants who are exposed to 4,000 images selected from the ImageNet dataset [22]. The dataset features image stimuli labeled at varying levels of granularity, including 40 images with coarse labels and 40 with fine-grained labels. This diverse labeling schema facilitates the study of both broad and specific neural representations associated with visual stimuli. The Things-EEG2 dataset [10] comprises EEG recordings from 10 participants, each exposed to 82,160 trials spanning 16,740 image conditions sourced from the THINGS database [13]. The Things-MEG dataset [12] includes MEG recordings from 4 participants, each exposed to 22,448 unique images covering 1,854 object categories. We follow the preprocessing pipelines used by the original papers of each dataset [7, 38]. EEG signals are re-referenced based on mastoid channels [37], and standard band-pass filtering and artifact removal techniques were applied to ensure high-quality data.

## 4.2 Implementation Details

The EEG-ImageNet dataset split for training and testing follows the setup outlined in the original paper. Specifically, the first 30 images of each category are used as the training set, and the last 20 images are allocated to the test set. For each subject, there are 1600 and 2400 images and corresponding brain signals in the test set and the training set, respectively. The Things-EEG and Things-MEG datasets follow the split outlined in the original paper, both with each participant generating 200 images during the test stage. The retrieval dataset is sampled from ImageNet21k [22], consisting of 8.5M randomly selected images used as the retrieval pool.

The input of brain encoder $S_i$ for EEG signals is differential entropy (DE) features extracted from the 40 ms to 440 ms time window [40, 41], which effectively captures the complexity and variability of brain activity in the frequency domain [9]. These features are calculated across five frequency bands: delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and gamma (30–80 Hz), resulting in a 310-dimensional input. For the MEG signals, we use the raw time-domain signals as feature inputs [3]. The brain encoder's output is designed to match the dimensionality of the CLIP ViT-L/14 [21] image encoder, producing a 768-dimensional representation.

The brain encoder adopt a multi-layer perceptron (MLP) architecture with two 512-dimensional linear layers. During training, InfoNCE loss is applied with a temperature parameter $\tau$ of 0.1 and a contrastive learning weight $\lambda$ of 0.2. The model is optimized using a learning rate of 0.01, with a linear warmup for the first 50 epochs followed by decay. The training is conducted over 500 epochs with a batch size of 128 and a dropout rate of 0.5. The I2I strength in the Diffusion Pipeline is set to 0.8 after the ablation study.

The Stable Diffusion Image Variations are used as the diffusion backbones (https://huggingface.co/lambdalabs/sd-image-variations-diffusers). This model is fine-tuned on top of SD1.4 [23]. We use PNDM scheduler for 20 denoising steps to generate the final image outputs. Each reconstructed image is produced at a resolution of 512 × 512 pixels.

## 4.3 Metrics

We evaluate the performance of our model using image quality metrics consistent with those employed in previous fMRI-based image reconstruction studies [20, 25, 26]. These metrics are categorized into low-level and high-level measures, reflecting both pixel-level and semantic-level properties of the reconstructed images. The low-level metrics include PixCorr, which measures the pixel-level correlation between the reconstructed and ground truth images, and SSIM (Structural Similarity Index) [35], which evaluates structural similarities based on luminance, contrast, and texture. Additionally, 2-way comparisons of Alex(2) and Alex(5) are used for comparisons, leveraging features extracted from the second and

fifth layers of AlexNet [15] to capture progressively higher-level visual features such as edges, textures, and basic shapes.

The high-level metrics of Inception and CLIP are also calculated based on a two-way comparison. The similarities of images are measured in the last pooling layer of InceptionV3 [30] and CLIP-vision [21], respectively. The 2-way comparison is performed following the approach of Ozcelik and VanRullen [20], where the metrics are calculated based on the similarity of the reconstructed image with the ground truth image as well as with k randomly selected images. The metric for each data sample is assigned as 1 or 0 when the reconstructed image is more similar or less similar to the ground truth image. Then the metrics are averaged across all pairwise comparisons for robust evaluation, with a chance performance of 0.5. We choose k = 500 for EEG-ImageNet, and k = 50 for Things-EEG and Things-MEG. The last two metrics, Eff and SwAV, use distance metrics derived from EfficientNet-B1 [32] and SwAV-ResNet50 [6], respectively, to evaluate high-level semantic coherence and clustering relationships. The above metrics provide a comprehensive assessment of reconstruction quality, capturing both fine-grained details and high-level semantic alignment, ensuring a robust evaluation of the model's performance.

When evaluating retrieval performance using NDCG, we calculate the relevance scores not in a point-wise manner (e.g., using binary 0/1 labels), but rather based on the cosine similarity between the embeddings of the brain signals and the retrieved images. This approach calculates relevance in a pair-wise manner, where the similarity between the brain embedding and each image embedding determines the relevance score for ranking the images.

## 4.4 Baselines

We adapt the approach of Takagi and Nishimoto [31] on our datasets as the basic baseline. This baseline method aligns brain signals with the CLIP embedding of corresponding images by mapping the neural features into the embedding space of CLIP. The resulting embeddings are then directly fed into Stable Diffusion 1.4, to generate reconstructed images. Further, we compare results reported by Li et al. [16] on Things-EEG, and compare results of Benchetrit et al. [3] and Li et al. [16] on Things-MEG.

# 5 Results

## 5.1 Overall Results

As shown in Table 1, our method outperforms the baseline across all key metrics on the EEG-ImageNet dataset. However, on the Things-EEG and Things-MEG datasets, our performance on low-level metrics is lower than ATM. This can be attributed to ATM incorporates large models that generate captions for images and align brain signals with text embeddings, which brings additional information to improve pixel-level similarity. Despite this, our method achieves higher scores on high-level metrics, including Inception, CLIP, and EffNet/SwAV distances, showing that the introduction of retrieval-augmented priors significantly enhances semantic consistency and coherence in the reconstructed images. These results underscore the effectiveness of integrating retrieved priors in the diffusion-based generation process, enabling our method to strike a balance between fine-grained details and high-level semantic alignment. Furthermore, when compared to fMRI-based methods, the inherent low signal-to-noise ratio (SNR) in EEG and MEG signals makes it challenging for our method to surpass fMRI in most metrics [14]. However, visual reconstruction with EEG and MEG signals provides a significant advantage for real-time and convenient applications and opens the door to practical implementations where continuous. It could be imaged that more paradigms could be feasible even if the current reconstruction quality is lower than that achieved with fMRI signals.

*5.1.1 Case analyses.* In addition to quantitative metrics, we present case studies in Figure 3 to showcase reconstruction results for subject 8 of the EEG-ImageNet dataset. From the examples, Figure 3(a) illustrates good cases where both retrieval and reconstruction effectively capture semantically related visual elements of the objects. For instance, in the examples of the electric locomotive and the brigantine, even when the retrieved images belong to incorrect categories or differ in details, such as orientation, the brain embedding effectively guides the diffusion process, enabling accurate reconstructions. These results demonstrate the robustness of our framework in leveraging neural control to produce semantically aligned outputs, even under weak retrieval priors. However, as seen in Figure 3(b), there remain limitations in accurately reconstructing low-level details of the objects, such as orientation, quantity, and color. For example, in the reconstruction of the pool table, the orientation differs from the ground truth, and in the examples of the capuchin and the daisy, the number of objects is inconsistent between the reconstruction and the ground truth. Additionally, in the grape example, while the reconstructed color of green grapes is realistic, it does not match the ground truth image's purple grapes. These challenges are likely influenced by the representational capacity of the CLIP embeddings, which may not encode detailed spatial or quantitative information. Furthermore, the limitations of the chosen diffusion backbone, which primarily focuses on generating semantically aligned but not necessarily spatially precise outputs, further contribute to these discrepancies.

*5.1.2 Relationship between retrieval performance and reconstruction performance.* Figure 4 illustrates the relationship between retrieval performance, measured by NDCG@50, and reconstruction semantic-level metrics (CLIP and Inception) for each subject. The results reveal a positive correlation, indicating that better retrieval performance leads to improved reconstruction quality. This trend highlights the critical role of the retrieval module in our framework, as high-quality priors provide stronger semantic guidance during the diffusion process. To further explore this relationship, we conducted additional experiments in Section 5.2, which demonstrate how retrieval datasets influence reconstruction quality under different configurations.

*5.1.3 Analyses of images with fine-grained labels.* Table 2 presents the semantic-level evaluation results for fine-grained category reconstruction, comparing the baseline with our proposed BReAD framework. The results show that BReAD achieves significant improvements across all evaluated metrics, including Inception and CLIP, which measure high-level semantic alignment, as well as Eff and SwAV, which assess structural and clustering consistency. These findings demonstrate that BReAD can generalize beyond strict label dependencies, effectively utilizing large-scale datasets

**Table 1: Quantitative assessments of image reconstruction methods using low-level and high-level metrics. Low-level metrics include PixCorr, measuring pixel-wise correlation, SSIM, evaluating structural similarity, and Alex2/Alex5, reflecting feature similarity at the second/fifth layers of AlexNet, respectively. High-level metrics include Inception, assessing semantic alignment, CLIP, evaluating multi-modal semantic similarity, and Eff/SwAV, capturing high-level semantic coherence through distance-based metrics. Higher values indicate better performance for metrics with ↑ and lower values are better for those with ↓. Missing values are from papers not reporting all metrics or metrics being nonapplicable.**

| Method | Dataset | Low-Level Metrics | | | | High-Level Metrics | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | PixCorr↑ | SSIM↑ | Alex2↑ | Alex5↑ | Inception↑ | CLIP↑ | Eff↓ | SwAV↓ |
| Takagi and Nishimoto [31] | EEG-ImageNet | 0.010 | 0.196 | 0.552 | 0.598 | 0.657 | 0.673 | 0.975 | 0.651 |
| **BReAD (ours)** | | **0.064** | **0.221** | **0.621** | **0.738** | **0.739** | **0.812** | **0.868** | **0.540** |
| Takagi and Nishimoto [31] | Things-EEG | 0.129 | 0.291 | 0.689 | 0.785 | 0.660 | 0.667 | 0.979 | 0.648 |
| NICE [29] | | 0.142 | 0.276 | 0.739 | 0.832 | 0.659 | 0.722 | - | 0.612 |
| ATM[16] | | **0.160** | **0.345** | **0.776** | 0.866 | 0.734 | 0.786 | - | 0.582 |
| **BReAD (ours)** | | 0.146 | 0.339 | 0.747 | **0.871** | **0.803** | **0.855** | **0.836** | **0.512** |
| Takagi and Nishimoto [31] (s1) | Things-MEG | 0.033 | 0.267 | 0.563 | 0.612 | 0.600 | 0.637 | 0.979 | 0.667 |
| Benchetrit et al. [3] | | 0.058 | 0.327 | **0.695** | **0.753** | 0.593 | 0.700 | - | 0.630 |
| ATM[16] | | **0.104** | **0.340** | 0.613 | 0.672 | 0.619 | 0.603 | - | 0.651 |
| **BReAD (ours, s1)** | | 0.067 | 0.298 | 0.619 | 0.734 | **0.643** | **0.737** | **0.911** | **0.585** |
| Brain-Diffuser [20] | NSD (fMRI) [1] | 0.254 | 0.356 | 0.942 | 0.962 | 0.872 | 0.915 | 0.775 | 0.423 |


**Figure 3: Reconstruction results sampled from subject S8 of the EEG-ImageNet dataset, including ground-truth images, retrieved priors, reconstructed images of BReAD, and reconstructed images of baseline [31]. (a) presents good cases where the reconstruction effectively captures the semantic content of the objects. (b) presents bad cases where low-level details of the objects, such as orientation, quantity, and color, still exhibit some flaws.**

with noisy or generic annotations to generate accurate and visually coherent reconstructions, even for fine-grained categories.

Case studies on images with fine-grained labels from the EEG-ImageNet dataset are presented in Figure 5. As shown in Figure 5,

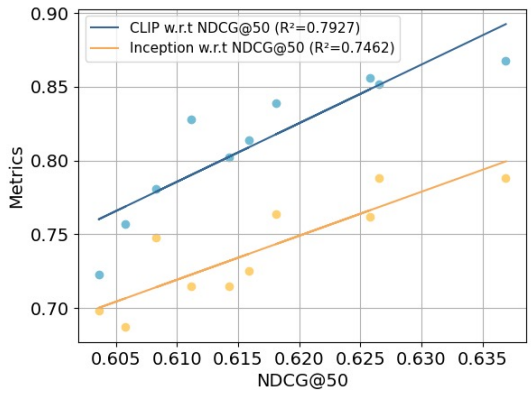


**Figure 4: The relationship between retrieval performance (measured by NDCG@50) and reconstruction semantic-level metrics (CLIP and Inception). Each dot represents a subject in the EEG-ImageNet dataset.**

**Table 2: Quantitative assessments of semantic-level metrics for fine-grained category reconstruction on the EEG-ImageNet dataset. Higher scores indicating better performance for ↑ metrics and lower scores for ↓ metrics. † indicates that the difference compared to the best-performing model is significant with p-value < 0.05.**

| Method | Inception↑ | CLIP↑ | Eff↓ | SwAV↓ |
|---|---|---|---|---|
| Takagi and Nishimoto [31] | 0.611 † | 0.598 † | 0.985 † | 0.662 † |
| **BReAD (ours)** | **0.694** | **0.727** | **0.915** | **0.570** |

the reconstructed images accurately reflect the fine-grained categories of the input brain signals, such as specific dog breeds, demonstrating the capability of the BReAD framework to preserve nuanced semantic information. The reconstructions successfully capture subtle distinctions between different breeds, such as Border Collies, Dobermans, and Samoyeds, validating the semantic fidelity of our method. Additionally, we analyzed the sources of the retrieved priors and found that many retrieved images did not originate strictly from categories corresponding to dog breeds. Instead, a significant portion came from abstract or generic labels in the ImageNet21k dataset, such as "adult" or "coal_black". This observation highlights the effectiveness of our BReAD framework in utilizing the diversity of the retrieval database. Even when the labels in the retrieval dataset are not perfectly aligned with the target categories, our method can extract and leverage relevant visual features from semantically related images, enabling robust reconstruction.

## 5.2 In-depth analysis

We further conduct an in-depth analysis to study the impact of the retrieval module on the image reconstruction performance. Table 3 presents the results analyzing the impact of img2img strength and retrieval dataset size on the performance of the image reconstruction pipeline.

**Effect of I2I Strength.** The I2I strength parameter controls the level of noise added to the retrieved priors during the forward diffusion process, effectively modulating the influence of the retrieved priors on the reconstructed images. As shown in Table 3, increasing the I2I strength leads to better performance on high-level metrics, indicating improved semantic alignment and coherence. However, low-level metrics decrease as the I2I strength increases, reflecting a loss of finer structural details. Conversely, reducing the I2I strength enhances the performance on low-level metrics, preserving pixel-level accuracy, but results in weaker high-level semantic alignment. This trade-off reflects the balance between leveraging the priors' structural guidance and allowing the generative process more flexibility to explore the latent space. It is important to note that setting the I2I strength to 1 would cause the generated image to almost entirely replicate the retrieved priors, losing control of the condition of the brain embeddings [20]. On the other hand, when I2I strength is set to 0, the pipeline effectively disables the retrieval module, resulting in reduced performance across most metrics, especially high-level ones, as it removes the guidance provided by the priors. This also highlights the effectiveness of our retrieval module, as incorporating semantically meaningful priors significantly enhances the quality of the reconstructed images.

**Effect of Retrieval Dataset Size.** The size of the retrieval dataset directly impacts the quality of the priors and, consequently, the reconstructed images. As shown in Table 3, increasing the retrieval dataset size from 50k to 8.5M significantly enhances high-level metrics, indicating better semantic consistency and feature diversity. This improvement can be attributed to the increased likelihood of retrieving images that are closely related to the true category of the original stimulus, providing richer semantic priors for the generative process. In contrast, smaller retrieval datasets often lack images that are semantically tied to the true category. As a result, the retrieval process tends to focus on other attributes, such as color distribution or composition, which are more readily available. This leads to slightly improved performance on low-level metrics, as these attributes are easier to match at a pixel level, even when semantic alignment is weak. The results also reveal that using retrieval datasets smaller than 50k introduces significant variability due to sampling randomness, rendering such configurations unreliable for meaningful evaluation. Due to computational resource constraints, we did not test datasets larger than 8.5M, though larger datasets could potentially provide even greater benefits for high-level metrics at the expense of increased resource requirements.

## 6 Conclusion

In this study, we proposed **BReAD** (**B**rain Image Reconstruction with **Re**trieval-**A**ugmented **D**iffusion), a novel framework for reconstructing visual images from brain signals. Our method integrates the strengths of retrieval-augmented models and diffusion-based generative processes to address the challenges in brain image reconstruction. Specifically, we introduced a brain encoder to map neural signals into a shared embedding space, enabling effective alignment with image representations. By incorporating a retrieval module, our framework leverages semantically relevant priors from


**Figure 5: Reconstruction results for fine-grained categories of "dog" in the EEG-ImageNet dataset, showing ground-truth images, retrieved priors, and reconstructed images.**

**Table 3: Analysis evaluating the impact of I2I strength and retrieval dataset size on the EEG-ImageNet dataset using low-level and high-level metrics. I2I strength refers to a parameter in the Diffusion Pipeline controlling the noise intensity added to the retrieved priors during forward diffusion, effectively modulating the extent to which the retrieved images influence the generation process. Higher I2I strength corresponds to stronger integration of retrieved priors into the generation. Retrieval dataset indicates the size of the dataset used for the Brain Image Retrieval stage, with 8.5M being the full retrieval dataset. Results for smaller datasets are obtained by randomly sampling subsets from the full retrieval dataset.**

| I2I strength | retrieval dataset | Low-Level Metrics | | | | High-Level Metrics | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | PixCorr↑ | SSIM↑ | Alex2↑ | Alex5↑ | Inception↑ | CLIP↑ | Eff↓ | SwAV↓ |
| 0.95 (w.o. condition) | 8.5M | 0.049 | 0.208 | 0.597 | 0.694 | **0.744** | **0.815** | 0.873 | 0.552 |
| 0.4 | 8.5M | 0.068 | 0.229 | 0.630 | 0.740 | 0.697 | 0.745 | 0.899 | 0.571 |
| 0 (w.o. retrieval) | - | **0.079** | 0.233 | **0.634** | 0.749 | 0.678 | 0.715 | 0.920 | 0.572 |
| 0.8 | 50k | 0.075 | **0.234** | 0.625 | **0.750** | 0.702 | 0.733 | 0.917 | 0.579 |
| 0.8 | 1M | 0.076 | 0.226 | 0.614 | 0.743 | 0.711 | 0.771 | 0.885 | 0.566 |
| 0.8 **(BReAD)** | 8.5M | 0.064 | 0.221 | 0.621 | 0.738 | 0.739 | 0.812 | **0.868** | **0.540** |

large-scale image datasets, which are refined into high-quality reconstructed images through a diffusion pipeline. This combination allows BReAD to balance semantic alignment and visual fidelity, addressing limitations in both low-level and high-level reconstruction quality observed in existing methods.

Through extensive experiments on several datasets, we demonstrated that BReAD outperforms baseline methods across a range of quantitative metrics, achieving superior performance in both pixel-level accuracy and semantic consistency. Qualitative analyses further validated the framework's ability to capture fine-grained details and generalize across diverse categories, even under noisy or abstract retrieval conditions. These results highlight the potential of BReAD as a robust approach for bridging neural signals and visual representations.

While our proposed BReAD framework demonstrates significant improvements in reconstructing visual images from brain signals, it has certain limitations that need further exploration. The brain encoder used in this work employs a relatively simple architecture, which may not fully capture the complexity and richness of neural signals. Additionally, due to computational resource constraints, we relied on a Diffusion backbone that, while effective, is not the most advanced model currently available in the field. Another notable observation is the consistent performance improvement with larger retrieval datasets. However, we did not explore whether this observation can hold with datasets larger than we used, leaving open the question of whether there exists an optimal dataset size for balancing computational cost and reconstruction quality. An important ethical limitation of our work is the reliance on brain signal data, which raises concerns about privacy and the potential misuse of sensitive neural information. While our study ensures the anonymization of data and adheres to ethical guidelines, future research must address privacy concerns more robustly, especially as we explore the scalability and real-world applications of brain-computer interfaces.

Future work can address these limitations by exploring deep network architectures specifically designed for processing neural signals, which could better model the intricate dynamics of brain activity and improve reconstruction fidelity. Additionally, cross-subject studies could be conducted to enhance the generalizability of the framework, enabling its application across diverse

individuals without requiring extensive subject-specific training data. Finally, the development of online applications capable of performing real-time reconstruction and feedback would significantly advance the practical utility of this research. Such implementations could pave the way for interactive brain-computer interfaces and other real-world applications, bridging the gap between research and deployment in fields like neuroscience, assistive technologies, and cognitive computing. Moreover, with sufficient computational resources, future work could explore end-to-end training of the diffusion components, specifically designing generative models tailored for brain signals. This would allow the diffusion model to be fully optimized for neural data, enhancing its ability to generate highly accurate visual reconstructions directly from brain activity.

## References


[1] Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. 2022. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience* 25, 1 (2022), 116–126.

[2] Yunpeng Bai, Xintao Wang, Yan-Pei Cao, Yixiao Ge, Chun Yuan, and Ying Shan. 2025. DreamDiffusion: High-Quality EEG-to-Image Generation with Temporal Masked Signal Modeling and CLIP Alignment. In *European Conference on Computer Vision*. Springer, 472–488.

[3] Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. 2023. Brain decoding: toward real-time reconstruction of visual perception. *arXiv preprint arXiv:2310.19812* (2023).

[4] Yanchao Bi. 2021. Dual coding of knowledge in the human brain. *Trends in Cognitive Sciences* 25, 10 (2021), 883–895.

[5] Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. 2022. Retrieval-augmented diffusion models. *Advances in Neural Information Processing Systems* 35 (2022), 15309–15324.

[6] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems* 33 (2020), 9912–9924.

[7] Alexandre D'Efossez, C. Caucheteux, J. Rapin, Ori Kabeli, and J. King. 2022. Decoding speech from non-invasive brain recordings. *ArXiv* abs/2208.12266 (2022).

[8] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* 34 (2021), 8780–8794.

[9] Ruo-Nan Duan, Jia-Yi Zhu, and Bao-Liang Lu. 2013. Differential entropy feature for EEG-based emotion classification. In *2013 6th international IEEE/EMBS conference on neural engineering (NER)*. IEEE, 81–84.

[10] Alessandro T Gifford, Kshitij Dwivedi, Gemma Roig, and Radoslaw M Cichy. 2022. A large and rich EEG dataset for modeling human visual object recognition. *NeuroImage* 264 (2022), 119754.

[11] Zijin Gu, Keith Jamison, Amy Kuceyeski, and Mert Sabuncu. 2022. Decoding natural image stimuli from fmri data with a surface-based convolutional network. *arXiv preprint arXiv:2212.02409* (2022).

[12] Martin N Hebart, Oliver Contier, Lina Teichmann, Adam H Rockter, Charles Y Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I Baker. 2023. THINGS-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *Elife* 12 (2023), e82580.

[13] Martin N Hebart, Adam H Dickter, Alexis Kidder, Wan Y Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I Baker. 2019. THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PloS one* 14, 10 (2019), e0223792.

[14] N Kannathal, U Rajendra Acharya, Choo Min Lim, and PK Sadasivan. 2005. Characterization of EEG—a comparative study. *Computer methods and Programs in Biomedicine* 80, 1 (2005), 17–23.

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25 (2012).

[16] Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, Haoyang Qin, and Quanying Liu. 2024. Visual decoding and reconstruction via eeg embeddings with guided diffusion. *arXiv preprint arXiv:2403.07721* (2024).

[17] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2021. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073* (2021).

[18] Rahul Mishra, Krishan Sharma, Ranjeet Ranjan Jha, and Arnav Bhavsar. 2023. NeuroGAN: image reconstruction from EEG signals via an attention-based GAN. *Neural Computing and Applications* 35, 12 (2023), 9181–9192.

[19] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[20] Furkan Ozcelik and Rufin VanRullen. 2023. Natural scene reconstruction from fMRI signals using generative latent diffusion. *Scientific Reports* 13, 1 (2023), 15666.

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[22] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. 2021. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972* (2021).

[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 234–241.

[25] Paul Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Aidan Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth Norman, et al. 2024. Reconstructing the mind's eye: fMRI-to-image with contrastive learning and diffusion priors. *Advances in Neural Information Processing Systems* 36 (2024).

[26] Paul S Scotti, Mihir Tripathy, Cesar Kadir Torrico Villanueva, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A Norman, et al. 2024. MindEye2: Shared-Subject Models Enable fMRI-To-Image With 1 Hour of Data. *arXiv preprint arXiv:2403.11207* (2024).

[27] Prajwal Singh, Dwip Dalal, Gautam Vashishtha, Krishna Miyapuram, and Shanmuganathan Raman. 2024. Learning Robust Deep Visual Representations from EEG Brain Recordings. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 7553–7562.

[28] Prajwal Singh, Pankaj Pandey, Krishna Miyapuram, and Shanmuganathan Raman. 2023. EEG2IMAGE: image reconstruction from EEG brain signals. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[29] Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. 2023. Decoding Natural Images from EEG for Object Recognition. *arXiv preprint arXiv:2308.13234* (2023).

[30] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2818–2826.

[31] Yu Takagi and Shinji Nishimoto. 2023. 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 (CVPR)*. 14453–14463.

[32] Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*. PMLR, 6105–6114.

[33] Chunyue Teng and Dwight J Kravitz. 2019. Visual working memory directly alters perception. *Nature human behaviour* 3, 8 (2019), 827–836.

[34] Michal Teplan et al. 2002. Fundamentals of EEG measurement. *Measurement science review* 2, 2 (2002), 1–11.

[35] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

[36] Dian Xie, Peiang Zhao, Jiarui Zhang, Kangqi Wei, Xiaobao Ni, and Jiong Xia. 2024. BrainRAM: Cross-Modality Retrieval-Augmented Image Reconstruction from Human Brain Activity. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 3994–4003.

[37] Dezhong Yao, Yun Qin, Shiang Hu, Li Dong, Maria L Bringas Vega, and Pedro A Valdés Sosa. 2019. Which reference should we use for EEG and ERP practice? *Brain topography* 32 (2019), 530–549.

[38] Ziyi Ye, Xiaohui Xie, Qingyao Ai, Yiqun Liu, Zhihong Wang, Weihang Su, and Min Zhang. 2024. Relevance Feedback with Brain Signals. *ACM Transactions on Information Systems* 42, 4 (2024), 1–37.

[39] Hong Zeng, Nianzhang Xia, Dongguan Qian, Motonobu Hattori, Chu Wang, and Wanzeng Kong. 2023. DM-RE2I: A framework based on diffusion model for the reconstruction from EEG to image. *Biomedical Signal Processing and Control* 86 (2023), 105125.

[40] Shaorun Zhang, Zhiyu He, Ziyi Ye, Peijie Sun, Qingyao Ai, Min Zhang, and Yiqun Liu. 2024. EEG-SVRec: An EEG Dataset with User Multidimensional Affective Engagement Labels in Short Video Recommendation. *arXiv preprint*

*arXiv:2404.01008* (2024).

[41] Shuqi Zhu, Ziyi Ye, Qingyao Ai, and Yiqun Liu. 2024. EEG-ImageNet: An Electroencephalogram Dataset and Benchmarks with Image Visual Stimuli of Multi-Granularity Labels. *arXiv preprint arXiv:2406.07151* (2024).