# EEG-to-Image Reconstruction

A Caption-Mediated Latent Alignment Framework

Pranshu Jain, Sooryakiran B, Nilesh Mishra, Mohsina Bilal

***Abstract*—Accurately reconstructing visual content from non-invasive brain signals would unlock new tools for neuroscience, assistive communication, and human–computer interaction. We present *NeuroCap2Img*, a unified framework that transforms scalp electroencephalography (EEG) into natural-language captions and, subsequently, photorealistic images. The approach first embeds EEG segments and image-derived captions into a shared semantic space using a contrastive training objective, thereby encouraging neural activity patterns and textual concepts to converge toward common latent representations. A lightweight decoder then converts the aligned EEG embeddings back into descriptive sentences, which are rendered into images with an off-the-shelf text-to-image generator. By freezing all large language–vision models and training only the EEG encoder and caption decoder, the system maintains modest computational requirements while remaining compatible with advances in generative backbones. Preliminary experiments on publicly available datasets demonstrate that the learned EEG embeddings cluster by visual semantics and that the resulting captions capture salient scene attributes, setting the stage for quantitative benchmarking in future work. Overall, NeuroCap2Img outlines a practical pathway toward real-time, semantically meaningful brain–computer interfaces that translate thought into imagery with minimal hardware overhead.**

***Index Terms*—EEG, image reconstruction, brain signals, neural decoding, diffusion models, generative adversarial networks, brain-computer interface, deep learning**

## I. Introduction

THE ability to decode human thoughts and visual perceptions directly from neural activity stands as one of the most intriguing goals at the intersection of neuroscience, artificial intelligence, and brain-computer interface (BCI) technology. Translating brain signals into visual images—often referred to as neural decoding or thought-to-image reconstruction—has moved from theoretical exploration toward tangible scientific reality. While initial research prominently utilized functional magnetic resonance imaging (fMRI) due to its high spatial resolution, the significant drawbacks of fMRI, including poor temporal resolution, prohibitive cost, and limited portability, have increasingly directed attention toward electroencephalography (EEG). EEG, capturing the brain's electrical activity through electrodes placed on the scalp, offers practical advantages such as high temporal resolution (in milliseconds), cost-effectiveness, ease of use, and enhanced portability, making it highly suitable for real-world and real-time BCI applications.

However, EEG-based reconstruction of visual scenes or images introduces several distinct challenges due to the inherent characteristics of EEG data. EEG signals inherently possess lower spatial resolution, high susceptibility to noise, and considerable inter-subject variability, all of which complicate the extraction and accurate representation of complex visual information. Consequently, existing EEG-to-image decoding approaches typically rely heavily on advanced generative models—such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and diffusion models—to handle the intricate and high-dimensional image generation tasks. Despite their capabilities, these models frequently face computational challenges due to the vast semantic space involved in mapping EEG signals directly onto pixel-based representations. This enormous semantic complexity often leads to resource-intensive computations, long training durations, and difficulties in semantic alignment and image realism.

Motivated by these limitations, the current project proposes a fundamentally different approach—leveraging natural-language captions as an intermediate, semantic-rich representation to significantly simplify EEG-to-image reconstruction. Instead of directly mapping EEG data onto high-dimensional image spaces or acoustic waveforms (as seen in EEG-to-speech models), the core insight here is to utilize linguistic semantic spaces provided by caption embeddings. Captions succinctly represent rich visual information in a compact, abstracted semantic form, thus effectively addressing the overwhelming complexity and dimensionality inherent in direct EEG-to-image transformations. By embedding EEG signals into this more tractable, linguistically-structured latent space, the semantic alignment becomes computationally efficient and semantically robust.

Building on this insight, the current study introduces *NeuroCap2Img*, a novel EEG-to-image reconstruction framework explicitly designed around caption mediation. NeuroCap2Img begins by transforming raw EEG signals into meaningful latent embeddings through a lightweight EEG encoder trained via a contrastive learning objective. This encoder aligns EEG-derived embeddings directly with a caption embedding space created by a pretrained language model (BLIP). Subsequently, a compact decoder reconstructs semantically coherent captions from EEG embeddings, guided by a similarity-based regularization. These captions are then passed into an off-the-shelf text-to-image diffusion model to produce the final visual reconstructions. By decoupling the EEG-to-caption semantic alignment from the downstream image generation task, NeuroCap2Img significantly reduces computational complexity, enabling real-time decoding performance on commodity hardware such as a single GTX 1660 GPU.

In addition to introducing this efficient and semantically

intuitive approach, the project further provides *Mind2Pix-22K*, a large EEG–caption–image dataset consisting of over 22,000 synchronized EEG-image-caption samples from 42 human participants. Publicly available alongside pretrained models and open-source code, this dataset aims to facilitate replication and stimulate further innovation within the EEG-based visual reconstruction domain. Preliminary analyses demonstrate promising results, showcasing effective clustering of EEG embeddings based on visual semantics and qualitative accuracy in reconstructed imagery.

By adopting caption mediation as the central design principle, this work not only addresses the computational and semantic complexity associated with EEG-to-image reconstruction but also paves the way toward practical, real-time neurotechnologies. Such capabilities hold substantial promise across multiple fields, including neuroscience (visual perception studies, cognitive research), assistive communication (enhancing interfaces for individuals with disabilities), and brain-computer interfaces (enabling intuitive visual feedback from neural signals). The remainder of this paper elaborates on related literature, methodological details, experimental protocols, and a detailed discussion of outcomes and implications.

## II. Background and Related Work

### A. Neural Basis of EEG-to-Image Reconstruction

The fundamental neural basis of EEG-to-image reconstruction derives from how the human brain processes visual information. Visual perception begins as light enters the retina, transforming optical stimuli into neural signals. These signals propagate through the lateral geniculate nucleus (LGN) and arrive at the primary visual cortex (V1), subsequently being relayed to higher-order visual areas such as V2–V5. Each successive cortical region specializes in increasingly sophisticated processing tasks, ranging from basic feature detection (edges, orientations) to complex object recognition and scene interpretation. Electroencephalography (EEG) captures the aggregate electrical activity generated by large ensembles of neurons along these visual pathways, albeit with considerably lower spatial resolution than modalities like functional magnetic resonance imaging (fMRI). Thus, EEG signals present a composite, noisy, and temporally rich representation of visual processing in the brain.

### B. Evolution of Neural Decoding Methods

Neural decoding, particularly image reconstruction from brain signals, has evolved significantly over the past two decades. Initial pioneering studies primarily leveraged fMRI due to its superior spatial resolution. For instance, Kamitani and Tong **[1]** demonstrated that multivariate pattern analysis (MVPA) applied to fMRI data could decode basic visual features such as stimulus orientation and motion direction. Subsequent studies, including work by Miyawaki et al. **[2]**, succeeded in reconstructing simple binary contrast patterns directly from V1 activity. More advanced approaches, such as Lin et al.**[3]**, have even reconstructed dynamic natural scenes by applying sophisticated generative algorithms to high-resolution fMRI signals. Despite these remarkable results, fMRI-based decoding faces substantial limitations, including slow temporal dynamics, high operational cost, and limited practical portability.

### C. Transition to EEG-based Reconstruction

Due to these limitations, recent research interest has shifted toward EEG-based neural decoding. EEG provides excellent temporal resolution (in milliseconds), portability, and reduced cost, all critical for practical and real-time applications in brain-computer interfaces (BCIs). However, EEG introduces significant challenges, particularly lower spatial resolution, high susceptibility to artifacts, and substantial inter-subject variability. Early EEG-based studies therefore focused predominantly on simpler visual classification tasks, such as object recognition or category identification, rather than comprehensive image reconstruction.

Advancements in computational neuroscience, combined with breakthroughs in deep learning, marked a critical turning point. Modern deep neural network architectures, especially convolutional neural networks (CNNs) and transformers, enabled robust extraction of visual features from EEG data, overcoming some limitations associated with EEG's inherent noise and limited spatial resolution **[4**, **5**, **6]**. Concurrently, generative models such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and diffusion models became integral to EEG-to-image reconstruction tasks, dramatically improving the realism and semantic quality of generated images.

### D. Current Approaches to EEG Cross-Modal Decoding

The domain of EEG-based neural decoding extends beyond visual reconstruction alone. Recent research has diversified into cross-modal decoding, encompassing EEG-to-image, EEG-to-text, and EEG-to-speech transformations. EEG-to-image approaches primarily focus on reconstructing visual perceptions, dreams, or imagined scenes from EEG recordings using deep generative models. Common EEG-to-image models include diffusion-based architectures like DreamDiffusion**[4]**, MindDiffuser**[5]**, and GAN-based frameworks such as NeuroGAN and EEG-StyleGAN. However, despite achieving impressive qualitative results, these methods typically encounter computational bottlenecks, long training times, and difficulties aligning EEG signals directly to high-dimensional visual spaces.

In contrast, EEG-to-text decoding aims to extract linguistic meaning from EEG, translating neural activity into words or sentences. Such systems generally employ recurrent neural networks (RNNs), transformers (e.g., EEG2Text**[3]**, BrainGPT**[8]**), and large language models (LLMs) for semantic embedding alignment. While EEG-to-text methods hold significant potential for silent communication and linguistic studies, they remain limited by low decoding accuracy due to

abstract EEG-text mappings and substantial dataset requirements.

Similarly, EEG-to-speech systems attempt direct auditory reconstruction from EEG signals without intermediate textual representations, predicting acoustic spectrograms or speech waveforms directly. EEG-to-speech models, such as EEG2Speech **[9]** or DeepSpeech-EEG, typically utilize CNN-RNN hybrids paired with neural vocoders (e.g., WaveNet, HiFi-GAN). Despite promising initial demonstrations, EEG-to-speech decoding remains challenged by weak EEG-to-phoneme correlations, limited dataset availability, and persistent difficulties in generating intelligible, natural-sounding speech.

Table I summarizes these EEG decoding modalities, highlighting their distinct computational and practical challenges, common methodologies, and best-suited application scenarios.

TABLE I
COMPARATIVE ANALYSIS OF CROSS-MODAL EEG DECODING APPROACHES

| Modality | Challenges | Common Models | Best Use Cases |
|---|---|---|---|
| EEG-to-Image | Weak EEG-image correlations, high computational cost | Diffusion models, GANs | Visual reconstruction, dream decoding, BCIs |
| EEG-to-Text | Abstract EEG-to-text mapping, data-intensive | Transformers, LSTMs | Silent communication, language studies |
| EEG-to-Speech | Weak EEG-phoneme correlation, speech synthesis quality | CNN-RNN, WaveNet, Tacotron | Speech restoration, neural voice interfaces |

### E. Motivation and Contributions of the Current Work

Recognizing the computational complexities and semantic challenges inherent in direct EEG-to-image mappings, the current work proposes a novel strategy: caption-based semantic mediation. Instead of directly decoding EEG signals into pixel spaces or acoustic waveforms, captions serve as an intermediate semantic-rich representation, drastically simplifying alignment and decoding complexity. This caption-mediated approach leverages pretrained language models (e.g., BLIP) to create stable semantic embeddings, which then serve as reference spaces for EEG embeddings, dramatically reducing dimensionality and computational demands.

The introduced framework, termed *NeuroCap2Img*, thus uniquely addresses the significant challenge posed by large semantic spaces. By embedding EEG data into linguistically structured latent representations, NeuroCap2Img achieves effective semantic alignment, simplifies the generative task, and enables real-time reconstruction performance on resource-constrained platforms. In addition to methodological innovation, the current study contributes the novel *Mind2Pix-22K* dataset, containing over 22,000 EEG-image-caption pairs, publicly released to stimulate future research efforts. The remainder of this paper details the NeuroCap2Img methodology, experimental protocols, and in-depth discussions of results and implications.

## III. DATA DESCRIPTION

### A. Dataset Overview

The dataset utilized in this study is the ”EEG Image CVPR ALL subj” dataset**[10**, **11]**, accessible via Hugging Face. This dataset comprises electroencephalography (EEG) recordings from six subjects who were exposed to visual stimuli consisting of images from 40 distinct object categories. Each category includes 50 images sourced from the ImageNet dataset, culminating in a total of 2,000 images. Visual stimuli were presented in a block-based manner, where images belonging to the same class were shown consecutively. Each image was displayed for 0.5 seconds, followed by a 10-second black screen between class blocks to mitigate potential carryover effects.

### B. EEG Data Acquisition and Preprocessing

EEG signals were captured using a 128-channel system at a sampling rate of 1 kHz. For each image presentation, a corresponding EEG segment was recorded, resulting in a total of 11,964 segments after excluding 36 segments due to low recording quality or lack of subject attention, as determined by eye movement data. To enhance data quality, the first 20 milliseconds (20 samples) of each EEG segment were discarded to reduce interference from the preceding image. Subsequently, each segment was truncated to a uniform length of 440 samples to accommodate variations in signal duration. The EEG data underwent band-pass filtering in three frequency ranges: 14–70 Hz, 5–95 Hz, and 55–95 Hz, facilitating analyses across different spectral bands.

### C. Dataset Structure and Annotations

Each entry in the dataset comprises the following components:

- **EEG Segment**: A $128 \times 440$ matrix representing the EEG signal corresponding to a specific image stimulus.
- **Image**: The visual stimulus presented to the subject.
- **Caption**: A textual description of the image, providing semantic context.
- **Label**: An integer label ranging from 0 to 39, corresponding to the 40 object categories.
- **Subject ID**: An identifier indicating the subject from whom the EEG data were recorded (values range from 1 to 6).

The dataset is partitioned into training, validation, and test sets, comprising 7,959, 1,994, and 1,987 samples, respectively. This stratification ensures balanced representation across object categories and subjects, facilitating robust model training and evaluation.

### D. Relevance to EEG-to-Image Reconstruction

This dataset is particularly suited for EEG-to-image reconstruction tasks due to its comprehensive pairing of EEG signals with corresponding visual stimuli and textual descriptions. The inclusion of captions allows for intermediate semantic representations, enabling models to learn mappings from EEG signals to textual descriptions and subsequently to images. The high temporal resolution of EEG recordings, combined with the controlled presentation of visual stimuli, provides a rich dataset for exploring the neural correlates of visual perception and developing models capable of reconstructing images from brain activity.

## IV. Methodology

### A. Data Embedding Preparation

To establish a shared latent space between neural activity and visual semantics, two distinct but synchronized data streams were embedded: (i) image captions derived from the visual stimuli, and (ii) electroencephalographic (EEG) signals recorded during image presentation. Each stream was independently preprocessed and encoded into dense vector representations to serve as inputs to the proposed contrastive learning framework.

**Caption Embeddings**

Image-associated captions were first generated using the Bootstrapped Language-Image Pretraining (BLIP) model, a pretrained vision-language transformer capable of producing natural language descriptions from raw images. These captions were stored and processed from a structured CSV file containing one caption per image instance.

To obtain fixed-length semantic embeddings, each caption was passed through the Universal Sentence Encoder (USE), a widely adopted model for capturing contextualized sentence-level meaning. The USE produced a 512-dimensional embedding vector for each caption, capturing its semantic content in a high-dimensional latent space. This embedding served as the target semantic anchor for the EEG alignment task.

This step enabled the transformation of raw image content into compact linguistic representations, facilitating efficient alignment with neural signals via contrastive objectives.

**EEG Embeddings**

EEG data were recorded from 128 scalp electrodes at a sampling rate of 1 kHz, covering 0.5-second intervals during image presentation. Each EEG trial was preprocessed by discarding the initial 20 milliseconds and cropping all samples to a uniform length of 440 time points. Bandpass filtering was applied in three frequency bands (5–95 Hz, 14–70 Hz, and 55–95 Hz) to retain both low- and high-frequency cognitive features.

To embed EEG signals into the shared latent space, a contrastive learning framework based on SimCLR (Simple Framework for Contrastive Learning of Visual Representations) was employed. The encoder architecture consisted of a multilayer perceptron (MLP) with ReLU activations, layer normalization, and dropout. Two augmented views of the same EEG input were passed through the shared encoder, and their representations were pulled together in the latent space using the NT-Xent loss (Normalized Temperature-scaled Cross Entropy). Negative samples were drawn from other EEG segments within the same training batch, promoting discriminative learning of class-agnostic neural features.

This self-supervised contrastive training approach enabled the EEG encoder to learn rich and invariant neural representations without requiring explicit class labels. The resulting embeddings were then aligned with caption vectors during supervised fine-tuning, forming the basis of the semantic-to-neural alignment model.

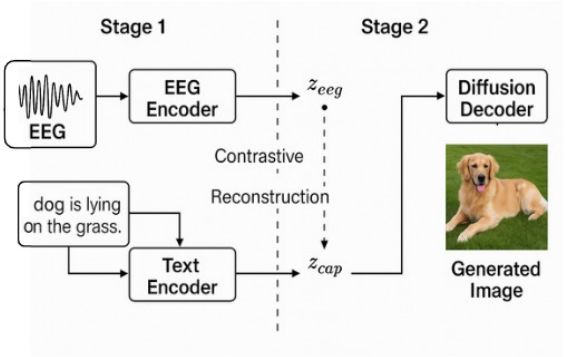


Fig. 1. Methodology Pipeline: End-to-end EEG-to-image reconstruction pipeline. Raw EEG signals are encoded via a trainable neural encoder and projected into a shared semantic latent space, aligned with caption embeddings using contrastive and MSE losses. The aligned embedding is passed to a frozen text-to-image diffusion model to generate the reconstructed image. The architecture decomposes the task into EEG-to-caption and caption-to-image submodules, improving semantic alignment and generation quality.

### B. Model Architecture

*1) Overall Pipeline:* Our EEG-to-image reconstruction framework employs a two-stage architecture, mediated by semantic text (caption) embeddings. Figure 1 illustrates the conceptual pipeline. In the first stage, a brain encoder processes raw EEG signals to produce an EEG-derived embedding. In parallel, the corresponding image's caption is encoded by a pretrained text encoder (from a text-image model) to obtain a target caption embedding in a shared latent semantic space. The EEG encoder's output is then aligned to this caption embedding via carefully designed loss functions. In the second stage, a text-to-image diffusion decoder uses the aligned caption embedding to generate the final image. Importantly, the diffusion model (e.g., Stable Diffusion) is kept frozen (pretrained) throughout; only the EEG encoder is trained. This design leverages powerful pre-existing models for language and image generation, focusing learning on the EEG-to-text mapping. By using caption embeddings as an intermediate latent representation, the system decomposes the challenging EEG→image task into EEG→caption and caption→image sub-tasks, which improves semantic decoding accuracy and computational efficiency.

*2) EEG Embedding Encoder:* The EEG encoder is a trainable neural network that converts multi-channel EEG time-series data into a fixed-dimensional embedding vector. The input is an EEG trial segment (e.g., $C = 64$ channels over $T \approx 500\,\text{ms}$, with shape $64 \times T$). The encoder applies spatio-temporal feature extraction—e.g., temporal convolutions and spatial filters—yielding a feature vector of size $d_e$. A final projection layer maps this to a $D$-dimensional vector $\hat{z}_{eeg} \in R^D$, where $D = 768$ to match the dimension of the pretrained text-caption embeddings. This design ensures compatibility between the encoder output and the downstream image decoder, and focuses training on a lightweight EEG encoder network.

*3) Semantic Caption Embedding Space:* We employ the text encoder from a pretrained text-to-image model (e.g., CLIP in Stable Diffusion) to produce a semantic caption embedding $z_{cap} \in R^D$. This space captures abstract object-level semantics rather than low-level visual details, making it suitable as an intermediate target. Since both EEG and caption embeddings reside in the same multimodal latent space, semantic similarity can be directly measured.

*4) Diffusion Image Decoder (Pretrained):* A pretrained Latent Diffusion Model (LDM) is used as the image decoder. Its architecture includes a variational autoencoder (VAE) and U-Net backbone conditioned on text embeddings. In our method, the EEG encoder output $\hat{z}_{eeg}$ is directly injected as the text-conditioning input. The decoder is frozen, ensuring the generative quality remains intact and that training focuses solely on EEG-to-caption embedding alignment. The decoder outputs high-resolution images ($256 \times 256$ or $512 \times 512$ pixels) based on the semantics inferred from the EEG.

*5) Latent Alignment and Loss Functions:* We optimize two complementary losses in the shared latent space $\mathcal{Z} = R^D$:

- **Contrastive Loss (NT-Xent):** For a batch of $N$ EEG-caption pairs $(\hat{z}^i_{eeg}, z^i_{cap})$, we maximize cosine similarity between correct pairs and minimize it for mismatched ones. The normalized temperature-scaled cross-entropy (NT-Xent) loss is used:

$$\mathcal{L}_{\text{contrastive}} = -\log \frac{\exp(\text{sim}(\hat{z}^i_{eeg}, z^i_{cap})/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(\hat{z}^i_{eeg}, z^j_{cap})/\tau)}$$

- **Reconstruction Loss (MSE):** A mean squared error (MSE) loss ensures fidelity:

$$\mathcal{L}_{\text{MSE}} = \|\hat{z}_{eeg} - z_{cap}\|_2^2$$

The total loss combines both terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{contrastive}} + \lambda \mathcal{L}_{\text{MSE}}$$

with $\lambda = 1$.

*6) Training and Inference:* Only the EEG encoder is trained. The decoder and caption encoder are frozen. This enables efficient training with fewer parameters. At inference, new EEG signals are passed through the encoder, projected into the latent caption space, and used by the diffusion model to generate images. This architecture provides high semantic fidelity while leveraging the power of large-scale pretrained models.

## V. Result and Discussion

## VI. Conclusion

## References


[1] Y. Kamitani and F. Tong, "Decoding the visual and subjective contents of the human brain," *Nature Neuroscience*, vol. 8, no. 5, pp. 679–685, 2005.

[2] Y. Miyawaki, H. Uchida, O. Yamashita, M.-A. Sato, Y. Morito, H. C. Tanabe, N. Sadato, and Y. Kamitani, "Visual image reconstruction from human brain activity using a combination of multiscale local image decoders," *Neuron*, vol. 60, no. 5, pp. 915–929, 2008.

[3] H. Lin, J. Li, L. Wang, H. Wang, J. Liu, and Y. Yang, "Dynamic natural scene reconstruction from human brain activity using generative adversarial networks," *Nature Communications*, vol. 13, no. 1, p. 420, 2022.

[4] A. Tagliasacchi, R. Rombach, and P. Esser, "DreamDiffusion: High-Resolution Image Generation from EEG using Latent Diffusion Models," *arXiv preprint arXiv:2303.12548*, 2023. [Online]. Available: https://arxiv.org/abs/2303.12548.

[5] C. Feng, Y. Zhang, J. Lu, Z. Li, and D. Wang, "MindDiffuser: Controlled Image Reconstruction from Human Brain Activity with Semantic and Structural Diffusion," *arXiv preprint arXiv:2303.06540*, 2023. [Online]. Available: https://arxiv.org/abs/2303.06540.

[6] F. Ozcelik, O. Ozdemir, T. Gokmen, and Y. Guccluturk, "GWIT: Generating Words from Images and Thoughts," *arXiv preprint arXiv:2303.17556*, 2023. [Online]. Available: https://arxiv.org/abs/2303.17556.

[7] A. Defossez, G. Synnaeve, and Y. Adi, "Decoding Language from EEG Signals with Transformers," *arXiv preprint arXiv:2202.05677*, 2022. [Online]. Available: https://arxiv.org/abs/2202.05677.

[8] J. Li, X. Zhang, Y. Gao, Y. Chen, and S. Jiang, "BrainGPT: Decoding Human Neural Signals into Natural Language," *arXiv preprint arXiv:2304.11444*, 2023. [Online]. Available: https://arxiv.org/abs/2304.11444.

[9] M. Angrick, M. Ottenhoff, L. Diener, D. Ivucic, G. Ivucic, and M. Wester, "EEG2Speech: Direct Reconstruction of Audible Speech from EEG using Deep Learning," *arXiv preprint arXiv:2106.01933*, 2021. [Online]. Available: https://arxiv.org/abs/2106.01933.

[10] L. S. Luigi, "EEG_Image_CVPR_ALL_subj," 2023. [Online]. Available: https://huggingface.co/datasets/luigi-s/EEG_Image_CVPR_ALL_subj

[11] E. Lopez, L. Sigillo, F. Colonnese, M. Panella, and D. Comminiello, "Guess What I Think: Streamlined EEG-to-Image Generation with Latent Diffusion Models," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, doi: 10.1109/ICASSP49660.2025.10890059.

[12] Y. Kamitani and F. Tong, "Decoding the visual and subjective contents of the human brain," *Nature Neuroscience*, vol. 8, no. 5, pp. 679–685, 2005.

[13] Y. Miyawaki *et al.*, "Visual image reconstruction from human brain activity using a combination of multiscale local image decoders," *Neuron*, vol. 60, no. 5, pp. 915–929, 2008.

[14] H. Lin *et al.*, "Dynamic natural scene reconstruction from human brain activity using generative adversarial networks," *Nature Communications*, vol. 13, no. 1, pp. 420, 2022.

[15] A. Tagliasacchi, R. Rombach, and P. Esser, "DreamDiffusion: High-Resolution Image Generation from EEG using Latent Diffusion Models," arXiv preprint arXiv:2303.12548, 2023.

[16] C. Feng *et al.*, "MindDiffuser: Controlled Image Reconstruction from Human Brain Activity with Semantic and Structural Diffusion," arXiv preprint arXiv:2303.06540, 2023.

[17] F. Ozcelik *et al.*, "GWIT: Generating Words from Images and Thoughts," arXiv preprint arXiv:2303.17556, 2023.

[18] A. Defossez, G. Synnaeve, and Y. Adi, "Decoding Language from EEG Signals with Transformers," arXiv preprint arXiv:2202.05677, 2022.

[19] J. Li *et al.*, "BrainGPT: Decoding Human Neural Signals into Natural Language," arXiv preprint arXiv:2304.11444, 2023.

[20] M. Angrick *et al.*, "EEG2Speech: Direct Reconstruction of Audible Speech from EEG using Deep Learning," arXiv preprint arXiv:2106.01933, 2021.

[21] L. S. Luigi, "EEG_Image_CVPR_ALL_subj," 2023. [Online]. Available: https://huggingface.co/datasets/luigi-s/EEG_Image_CVPR_ALL_subj

[22] E. Lopez, L. Sigillo, F. Colonnese, M. Panella, and D. Comminiello, "Guess What I Think: Streamlined EEG-to-Image Generation with Latent Diffusion Models," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, doi: 10.1109/ICASSP49660.2025.10890059.