ESTJ-GD: Visual Reconstruction via EEG-Image Joint Space Learning with Guided Diffusion

Yanshu Zhoumen 31520241154541
Siyue Liang 31520241154489
Chenyu Hu 31520241154503
Longhao Liu 31520241154515

Abstract

Reconstructing human-perceived images from neural signals bridges computer vision and neuroscience but remains challenging. EEG-based methods often struggle with spatial discrepancies between EEG signals and visual data, hindering accurate classification and compromising image quality. This work proposes a novel framework to decode semantic information and primary image representations from EEG signals and reconstruct images using a pre-trained diffusion model. Our method encodes EEG signals by assigning distinct weights to brain regions and aligns them with image embeddings through ternary and similarity loss in high-dimensional space, fine-tuned via a classification task to extract semantic information. Low-level visual features are decoded using a three-stage EEG-image co-training strategy, addressing EEG data scarcity. Finally, a pre-trained diffusion model synthesizes high-quality visual reconstructions aligned with human perception. Experimental results highlight the framework's competitive performance in classification, retrieval, and reconstruction tasks, showcasing the potential of EEG-based decoding for brain-computer interfaces due to its portability, low cost, and high temporal resolution.

1 Introduction

Multimodal learning has recently gained attention, with text-to-image synthesis(Ramesh et al. 2021) and image-text contrastive learning(Jia et al. 2021) driving significant advancements. EEG-to-image reconstruction, a growing area of interest, involves collecting electroencephalography (EEG) data while subjects view images and reconstructing perceived images from these signals. In this work, we propose a diffusion model, transforming EEG signals into embeddings to guide image reconstruction. Challenges arise from EEG's low signal-to-noise ratio and the complexity of visual representations. Inspired by CLIP(Radford et al. 2021), which aligns text and image embeddings using a self-supervised contrastive framework, we adopt contrastive learning to align EEG and image embeddings, enabling the extraction of semantic information from EEG signals.

The human visual system, crucial for perceiving the environment and acquiring external information, continuously processes visual input, from basic patterns to complex scenes(Marr and Vaina 1982). Neuroscience aims to uncover how this processing occurs at the neural level through the study of neural coding. Brain-computer interfaces (BCIs) hold transformative potential, from enhancing human-machine interaction to aiding paralyzed patients. A major challenge is decoding and reconstructing human-perceived visual information using non-invasive brain recordings. Visual decoding and reconstruction not only deepen our understanding of brain processing but also drive advancements in BCI applications.

As a non-invasive technology, EEG has become a crucial tool in visual decoding research, bridging brain activity and the external world, with applications in clinical diagnosis(Sakkalis 2011) and brain-computer interfaces(Vaid, Singh, and Kaur 2015). Significant progress has been made in EEG-based visual decoding in recent years. Recent studies on visual stimulus reconstruction increasingly leverage deep generative models, particularly denoising diffusion models(Rombach et al. 2022). We propose using an EEG decoder to map EEG signals into embeddings that guide diffusion models in image generation. However, existing methods(Zeng et al. 2023a) often prioritize reconstructing semantic information while neglecting low-level visual features, such as color and texture. To address this, we aim to extract both high-level and low-level visual features from EEG. Given that brain regions like V1–V4 in the occipital lobe process low- to mid-level visual information, while the inferior temporal cortex processes high-level features, channels near specific regions may better decode relevant features. Thus, the model should integrate global brain information while focusing on specific areas to decode diverse EEG features effectively.

In EEG feature extraction, contrastive learning with images is commonly used to align the latent representations of EEG and image embedding spaces(Song et al. 2023). However, most contrastive learning methods, limited by the self-supervised framework, fail to achieve category-level alignment. Specifically, EEG embeddings focus on increasing the distance from mismatched images, which, while effective in separating different categories, unnecessarily increases the distance from non-identical images within the same category. To address this, we propose incorporating supervised learning into the alignment process. This approach reduces the distance between EEG embeddings and images of the same category while increasing the distance from different categories. By identifying shared EEG features within a cat-

egory, our method facilitates more effective stimulus reconstruction at the category level.

Our **contributions** are as follows:

- We introduce a diffusion model for image reconstruction, guided by two conditions. The primary condition is class guidance based on semantic information derived from EEG signals. The additional condition, provided by T2Iadapter-generated adapters, extracts image-related features from EEG inputs, such as color, depth, and texture, to refine reconstruction quality.
- We propose a novel neuro-attention mechanism that enables the model to focus on neural information from specific brain regions. This mechanism captures high-level semantic information from the temporal lobe and low-level visual details, such as color, from the occipital lobe.
- We develop a supervised embedding alignment approach that pushes EEG embeddings away from images of different categories while pulling them closer to images of the same category in the embedding space. This method enhances the model's ability to identify shared features within the same category, improving stimulus reconstruction accuracy.

2 Related Work

Neural encoding and decoding of visual information has been a longstanding focus in neuroscience and computer science. Functional magnetic resonance imaging (fMRI) has been widely employed to decode semantic information in visual processing. However, its non-portability, high cost, and operational complexity render it unsuitable for meeting the high-speed and practical demands of brain-computer interfaces (BCIs). In contrast, electroencephalography (EEG) offers advantages such as portability and high temporal resolution, making it an essential tool for BCI applications. Consequently, reconstructing images from EEG signals has rapidly gained traction in recent years.

For example, Spampinato et al.(Spampinato et al. 2017) used a deep convolutional generative adversarial network (DCGAN) to reconstruct visual stimuli by extracting semantic features from EEG signals. Tirupattur et al.(Tirupattur et al. 2018) proposed the ThoughtViz framework, which applied a conditional GAN to transform encoded EEG signals into corresponding images. While generative adversarial networks (GANs) and variational autoencoders (VAEs) have achieved some success in visual reconstruction, challenges persist in generating high-quality images due to EEG's low signal-to-noise ratio and significant inter-subject variability.

To address these limitations, Zeng et al.(Zeng et al. 2023a) introduced the EG-DDPM module, which leverages features extracted from EEG as guidance for a diffusion model to generate images. Additionally, Song et al.(Song et al. 2023) incorporated contrastive learning, employing image and EEG encoders to extract features from paired EEG-image data, further enhancing decoding performance.

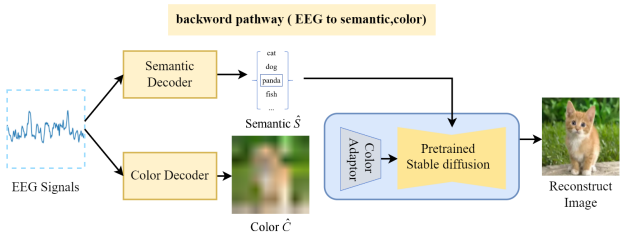

Figure 1: ESTJ-GD's Overview. Decoding semantic and colour details from EEG, which consists of three processes: 1) decoding semantics from EEG signals, 2) decoding colours from EEG signals, and 3)finally reconstructing the image completely by a pre-trained SD model.

3 Method

3.1 Problem Statement

The objective of visual decoding is to reconstruct an observed image $I \in \mathbb{R}^{H\times W\times 3}$ from brain activity signals elicited by visual stimuli. Electroencephalography (EEG) is typically employed to record these brain activities, representing them as a multivariate time series EEG $\in \mathbb{R}^{C\times S}$, where C is the number of channels and S is the number of time steps. Formally, the task aims to optimize a function $f(\cdot)$ such that $f(\text{EEG}) = \hat{I}$, where $\hat{I}$ closely approximates the original image I.

3.2 Overview of the Work

To tackle this task, we propose a framework, **ESTJ-GD**, inspired by the fundamentals of human perception. Our approach explicitly designs reverse visual pathways to decode the semantic and color information embedded in EEG data. Given the high noise and low resolution of EEG signals compared to fMRI, the extracted information is relatively limited. To address this, we leverage a pre-trained model to refine the reconstructed image after extracting the basic features.

Fig. 1 outlines the proposed model, which comprises three stages: Joint Spatial-Semantic Alignment, Three-Stage Decoding of Color, and Guided Image Reconstruction. These stages decompose the reverse mapping from EEG to image into three subprocesses: EEG $\rightarrow \{\hat{S}\}$, EEG $\rightarrow \{\hat{C}\}$, and $\{\hat{S}, \hat{C}\} \rightarrow I$.

In the first stage, semantic details are decoded from EEG through a reverse pathway. We employ joint attention coding across multiple brain regions (Sec. 3.3) and align these details with the semantic space of the image using ternary loss(Schroff, Kalenichenko, and Philbin 2015) and CLIP space embedding(Radford et al. 2021). To mitigate cross-domain issues in EEG datasets, we combine metric-based and classification-based approaches, overcoming the limitations of direct classification methods(Jiang, Fares, and Zhong 2020) that rely on consistent test set distributions.

In the second stage (Sec. 3.4), we exploit the correlation between semantics and color to decode color information from embeddings derived in the first stage. Both stages use joint spatial learning to align EEG and image embeddings

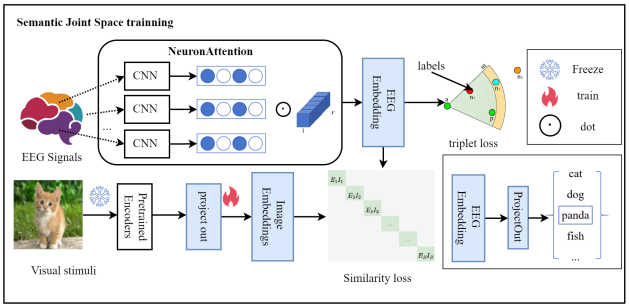

Figure 2: Overview of EEGTCLIP. EEG and image inputs are encoded into embeddings E_{eeg} and E_{img} through a neural encoder and a pre-trained image encoder, respectively. Following the CLIP methodology, we learn a joint representation of E_{eeg} and E_{img}. A triplet loss is then applied within each batch to align different E_{eeg} embeddings with their respective category spaces. Finally, triplet loss alignment is applied between E_{eeg} and the category spaces after the joint embeddings of E_{eeg} are fine-tuned for specific classification tasks.

in a high-dimensional space, improving decoding accuracy and model generalization.

Given the lossy nature of EEG data and the one-way transformation from image to EEG, we treat the decoding process as a generative task. Using extracted $\hat{S}$ and $\hat{C}$ as guidance, the third stage (Sec. 3.5) employs guided image reconstruction. Here, we follow recent visual decoding methods(Takagi and Nishimoto 2023; **?**) and utilize a frozen SD model with T2I-Adapter(Mou et al. 2024) to generate images guided by $\hat{S}$ and $\hat{C}$.

3.3 Semantic Decoder

Inspired by advanced time-series models, we propose a joint EEG-image space learning method, **EEGTriClip**, which aligns raw EEG signals to their feature space and refines them for signal-image classification. **Fig. 2** illustrates the semantic decoding phase workflow.

For EEG encoding, we divide the brain into five regions based on electrode positions. Each region is encoded using a 4-layer 1D CNN, with a learnable attention weight w_i assigned to each region. The region embedding e_i is calculated as $e_i = \mathcal{F}_i(E_i) \times w_i$, where $\mathcal{F}_i$ is the CNN encoder. The overall EEG encoding is $E_{\text{eeg}} = \sum_{i=1}^{r} e_i$. For image encoding, a pre-trained Swin Transformer is used, with its output mapped to the EEG space through a trainable projection layer, yielding the image embedding E_{img}.

In joint space learning, we begin with triplet loss on the EEG-Label space using semi-hard triplets to enhance feature differentiation. The triplet loss function is defined as:

$$\|f_\theta(x_a) - f_\theta(x_p)\| < \|f_\theta(x_a) - f_\theta(x_n)\| < \|f_\theta(x_a) - f_\theta(x_p)\| + \delta,$$

where x_a, x_p, and x_n are the anchor, positive, and negative samples, respectively, and δ is the margin. The model is then optimized with two loss functions:

$$\mathcal{L}_{\text{i2e}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(E^i_{\text{img}}, E^i_{\text{eeg}})/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(E^i_{\text{img}}, E^j_{\text{eeg}})/\tau)},$$

and

$$\mathcal{L}_{\text{e2i}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(E^i_{\text{eeg}}, E^i_{\text{img}})/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(E^i_{\text{eeg}}, E^j_{\text{img}})/\tau)}.$$

The combined loss is:

$$\mathcal{L} = \alpha\mathcal{L}_{\text{e2i}} + (1-\alpha)\mathcal{L}_{\text{i2e}},$$

where sim denotes cosine similarity, τ is the temperature, and α is the weighting coefficient. Finally, we fine-tune the EEG embeddings on specific semantic classification tasks.

3.4 Color Decoder

The goal of color decoding is to generate low-level images for background signal generation by the Diffusion model. To mitigate dataset limitations, we employ a three-stage encoding-decoding process involving the I2E-Encoder and E2I-Decoder, as shown in **Fig. 3**.

Stage 1 Given a set of pairs $(E, I) = \{\text{EEG, Image}\}$, we first train an encoder to map images to their corresponding EEG data. The encoder uses a 4-layer CNN with a residual structure, optimized with a combination of MSE and cosine proximity losses:

$$\mathcal{L}_r(E_{\text{eeg}}, \hat{E}_{\text{eeg}}) = \beta \cdot \text{MSE}(E_{\text{eeg}}, \hat{E}_{\text{eeg}}) - (1-\beta)\cos(\angle(E_{\text{eeg}}, \hat{E}_{\text{eeg}})),$$

where β is an empirically determined hyperparameter. The I2E-Encoder is trained through the joint space learning process outlined in Sec. 3.3.

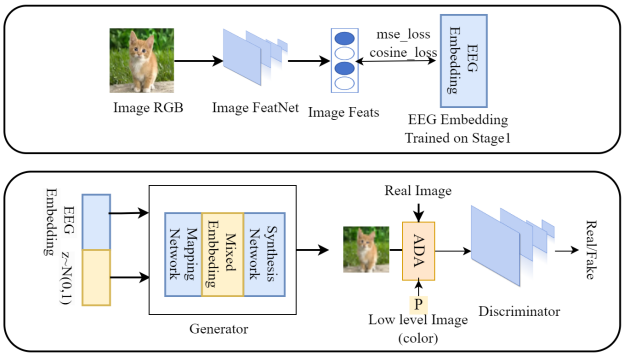

Figure 3: Overview of I2E-Encoder and E2I-Decoder: The encoder aligns the semantic spaces of EEG and image. A 4-layer CNN encodes the image, and the decoder generates low-level images from the EEG signal. The decoder uses ADA-StyleGAN to improve image quality and stability.

Stage 2 The decoding process is framed as an initial generation task using the StyleGAN-ADA model (Karras et al. 2020). This model synthesizes images by taking a feature vector and noise from an isotropic Gaussian distribution. StyleGAN-ADA enhances the discriminator's ability

to learn from limited data by augmenting real images during training. However, the GAN's performance is constrained by the limited dataset.

Stage 3 To improve generalization and address the scarcity of EEG data, we further train the decoder using a visual-to-decoding approach. The I2E-Encoder is frozen, and the image embeddings are mapped to the EEG feature space. These embeddings are then used as inputs for the E2I-Decoder to generate new images, allowing the model to capture robust visual representations without direct EEG data.

3.5 Guided Image Reconstruction

Following the training, we have identified the semantic category $\hat{S}$ associated with the EEG signal and generated its corresponding sketch $\hat{C}$. Using this information, we can reverse-engineer the visual process to infer the content. To reconstruct the final image from the EEG data, we employ Stable Diffusion (SD) (Li et al. 2024), with guidance provided by the color adapter $\mathcal{R}_c$ within the T2I-adapter (Mou et al. 2024). This process is formulated as $\hat{I} = \text{SD}(z, \mathcal{R}_c, \hat{S})$, where z is random noise.

4 Experiments

4.1 Experimental Settings

We have used EEGCVPR40 Dataset (Spampinato et al. 2017) for training and testing the EEG representation learning. The dataset consists of EEG-image pairs across 40 categories, with the images being a subset of the ImageNet dataset. During EEG recording, six subjects were shown 50 images per category within a 0.5-second window. After preprocessing, the EEG signals are represented with 128 channels and 440 time steps.

4.2 EEG Decoding Performance

To validate the effectiveness of our EEG semantic decoder, we conducted an image classification task. The classification accuracy, shown in Table 1 (a), reached 95.28%, significantly outperforming classical methods. This demonstrates our model's ability to effectively decode the semantic content of EEG signals.

Additionally, to assess the discriminative power of the learned features, we evaluated the k-means score. Our model achieved a score of 0.885, surpassing other methods, as seen in Table 1 (a). To further visualize the model's feature extraction ability, we employed t-Distributed Stochastic Neighbor Embedding (T-SNE) to project the features into a 2D space. As shown in Figure 4, the inter-class feature distribution expands, while intra-class features become more compact during the feature extraction process.

4.3 Image Generation Performance

Figure 5 qualitatively demonstrates our model's image generation performance. The synthetic images produced by our framework exhibit both diversity and high fidelity. To quantitatively evaluate the generation and reconstruction accuracy, we assess the images from both high-level and low-level perspectives. For low-level features, we use the Structural Similarity Index (SSIM) to measure pixel, structural,

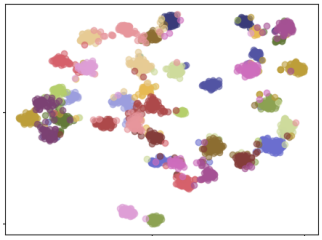

Figure 4: t-SNE Visualization of extracted features

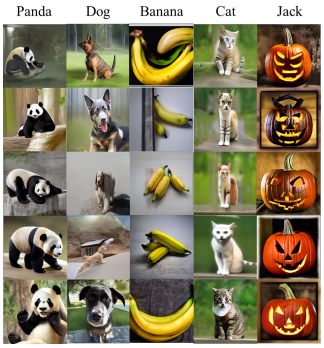

Figure 5: EEG to Image

and textural similarity, and Color Discrepancy (CD) to evaluate color consistency. Our model achieved scores of 0.335 for SSIM and 3.860 for CD, confirming its effectiveness in reconstructing low-level features like color.

For high-level features, we use the Inception Score (IS) and Fréchet Inception Distance (FID) (**?**)o provide deeper insights into the quality and diversity of the generated images. As shown in Table 1 [b, c], our model demonstrates strong performance across both metrics.

4.4 Ablation Study

Figure 6 presents the ablation study, focusing on three key aspects: (1) the effect of the reverse pathways for se-

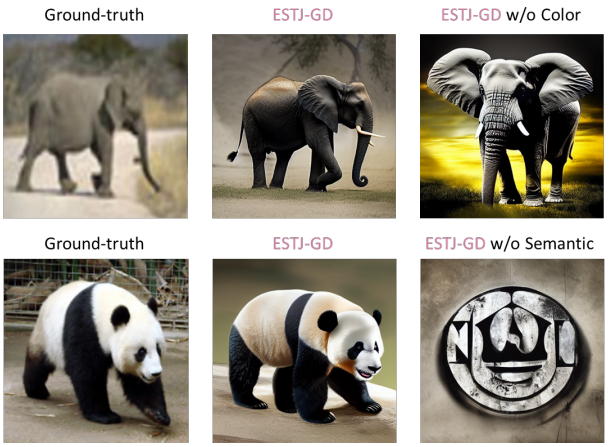

Figure 6: Effect of Semantic and Color Decoding

Table 1: Performance Comparison on Different Metrics

(a) Accuracy and K-means Performance

Method	Accuracy (%) ↑	K-means ↑
LSTM	83.36	0.450
EEGNet (Lawhern et al. 2018)	88.13	-
SyncNet (Li et al. 2017)	83.45	-
EEG-ChannelNet (Palazzo et al. 2020)	98.35	-
A-Bi-LSTM	94.15	-
EV-Net (Zeng et al. 2023c)	**98.98**	-
Neuro Vision (Khare et al. 2022)	98.8	-
ours	95.28	**0.885**

(b) Inception Score and FID Performance

Method	Inception Score ↑	FID ↓
DCVAE (Kavasidis et al. 2017)	1.98	-
DM-RE2I (Zeng et al. 2023b)	7.46	-
DCLS-GAN (Fares, Zhong, and Jiang 2020)	6.64	-
Brain2Image-VAE (Kavasidis et al. 2017)	4.49	-
NeuroVision	5.15	-
Improved-SNGAN (Zheng et al. 2020)	5.53	-
EEGStyleGAN-ADA (Singh et al. 2024)	**10.82**	174.13
ours	8.78	**165.35**

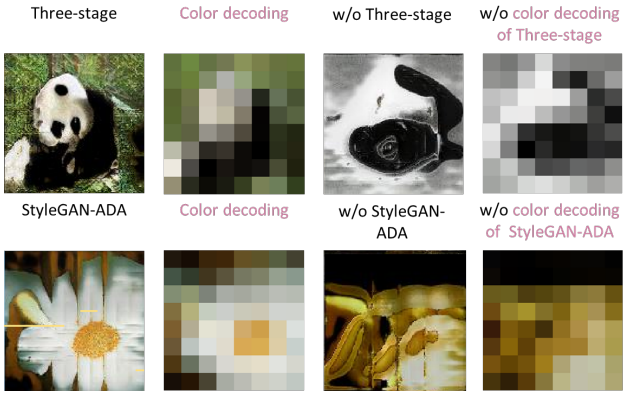

Figure 7: Effect of Three-stage Process and StyleGAN-ADA

mantic and color information decoded from EEG as guidance for image reconstruction, (2) the effectiveness of the Neuro-Attention mechanism in the semantic decoder, and (3) the role of the three-stage coding-decoding process and StyleGAN-ADA in color decoding.

Effect of Semantic Decoding. As shown in Figure 6, semantic decoding significantly enhances image reconstruction by providing essential semantic guidance, allowing the model to convey image meaning effectively. Without EEG-derived semantic guidance, reconstructed images fail to capture meaningful content, appearing random despite resembling real images in low-level features.

Effect of Color Decoding. As depicted in Figure 6, color decoding substantially improves visual quality by ensuring pixel-level similarity. Without EEG color guidance, reconstructed images fail to match the real images' color features, resulting in significant pixel-level discrepancies.

Effect of Neuro-Attention Mechanism. We conducted a comparative experiment to evaluate the Neuro-Attention mechanism. Without it, the EEG decoder achieved a classification accuracy of 85.71% and a k-means score of 0.652. With Neuro-Attention, these metrics improved significantly, demonstrating its ability to extract and utilize discriminative features from EEG signals for enhanced classification.

Effect of Three-stage Coding and Decoding Process. To address EEG data scarcity and improve generalization to unseen classes, we implemented a three-stage coding and decoding process using self-supervised learning. Figure 7 shows that this process preserves pixel-level similarity and meets basic color decoding requirements. Without it, the decoder loses critical color information, impairing the image reconstruction.

Effect of StyleGAN-ADA. StyleGAN-ADA, used as the decoder in the three-stage process, significantly outperforms CNN-based decoders. As shown in the lower part of Figure 7, StyleGAN-ADA preserves color features well, resulting in accurate and rich color information. In contrast, CNN-based decoders fail to retain key features, leading to poor color decoding.

5 Discussion and Conclusion

In this study, we developed the EEG-based image decoding and reconstruction framework, ESTJ-GD, which operates in three phases. First, EEG signals are encoded using joint spatial learning to align with both semantic and categorical image information. In the second phase, the generator is trained with an adversarial approach and a three-stage process (image-to-EEG, EEG-to-image, image-to-image), addressing data limitations. The final phase integrates a pre-trained stable diffusion model with a T2I-Adapter, using EEG-derived semantic and color information as conditions to guide image generation.

Diffusion models in EEG-based visual reconstruction tackle two challenges: (1) predicting the image's categorical identity and (2) generating visual sketches for refinement. Our model achieves 94% classification accuracy on unseen data, making it suitable for BCI devices. Using GANs as decoders enhances detail and supports accurate color generation. This method shows promise for generating visually and semantically consistent images in EEG-based applications.

However, performance varies between cross-subject and within-subject settings, likely due to brain heterogeneity and noise. Moreover, the method is not fully end-to-end, requiring multiple models for task completion. Future work will focus on cross-domain learning to reduce variability and simplify the process.

References

Fares, A.; Zhong, S.-h.; and Jiang, J. 2020. Brain-media: A dual conditioned and lateralization supported gan (dcls-gan) towards visualization of image-evoked brain activities. In *Proceedings of the 28th ACM International Conference on Multimedia*, 1764–1772.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, 4904–4916. PMLR.

Jiang, J.; Fares, A.; and Zhong, S.-H. 2020. A brain-media deep framework towards seeing imaginations inside brains. *IEEE Transactions on Multimedia*, 23: 1454–1465.

Karras, T.; Aittala, M.; Hellsten, J.; Laine, S.; Lehtinen, J.; and Aila, T. 2020. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33: 12104–12114.

Kavasidis, I.; Palazzo, S.; Spampinato, C.; Giordano, D.; and Shah, M. 2017. Brain2image: Converting brain signals into images. In *Proceedings of the 25th ACM international conference on Multimedia*, 1809–1817.

Khare, S.; Choubey, R. N.; Amar, L.; and Udutalapalli, V. 2022. Neurovision: perceived image regeneration using cprogan. *Neural Computing and Applications*, 34(8): 5979–5991.

Lawhern, V. J.; Solon, A. J.; Waytowich, N. R.; Gordon, S. M.; Hung, C. P.; and Lance, B. J. 2018. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of neural engineering*, 15(5): 056013.

Li, D.; Wei, C.; Li, S.; Zou, J.; Qin, H.; and Liu, Q. 2024. Visual decoding and reconstruction via eeg embeddings with guided diffusion. *arXiv preprint arXiv:2403.07721*.

Li, Y.; Dzirasa, K.; Carin, L.; Carlson, D. E.; et al. 2017. Targeting EEG/LFP synchrony with neural nets. *Advances in neural information processing systems*, 30.

Marr, D.; and Vaina, L. 1982. Representation and recognition of the movements of shapes. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 214(1197): 501–524.

Mou, C.; Wang, X.; Xie, L.; Wu, Y.; Zhang, J.; Qi, Z.; and Shan, Y. 2024. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4296–4304.

Palazzo, S.; Spampinato, C.; Kavasidis, I.; Giordano, D.; Schmidt, J.; and Shah, M. 2020. Decoding brain representations by multimodal learning of neural activity and visual features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11): 3833–3849.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, 8821–8831. Pmlr.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Sakkalis, V. 2011. Applied strategies towards EEG/MEG biomarker identification in clinical and cognitive research. *Biomarkers in medicine*, 5(1): 93–105.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 815–823.

Singh, P.; Dalal, D.; Vashishtha, G.; Miyapuram, K.; and Raman, S. 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.

Song, Y.; Liu, B.; Li, X.; Shi, N.; Wang, Y.; and Gao, X. 2023. Decoding Natural Images from EEG for Object Recognition. *arXiv preprint arXiv:2308.13234*.

Spampinato, C.; Palazzo, S.; Kavasidis, I.; Giordano, D.; Souly, N.; and Shah, M. 2017. Deep learning human mind for automated visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6809–6817.

Takagi, Y.; and Nishimoto, S. 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*, 14453–14463.

Tirupattur, P.; Rawat, Y. S.; Spampinato, C.; and Shah, M. 2018. Thoughtviz: Visualizing human thoughts using generative adversarial network. In *Proceedings of the 26th ACM international conference on Multimedia*, 950–958.

Vaid, S.; Singh, P.; and Kaur, C. 2015. EEG signal analysis for BCI interface: A review. In *2015 fifth international conference on advanced computing & communication technologies*, 143–147. IEEE.

Zeng, H.; Xia, N.; Qian, D.; Hattori, M.; Wang, C.; and Kong, W. 2023a. DM-RE2I: A framework based on diffusion model for the reconstruction from EEG to image. *Biomedical Signal Processing and Control*, 86: 105125.

Zeng, H.; Xia, N.; Qian, D.; Hattori, M.; Wang, C.; and Kong, W. 2023b. DM-RE2I: A framework based on diffusion model for the reconstruction from EEG to image. *Biomedical Signal Processing and Control*, 86: 105125.

Zeng, H.; Xia, N.; Tao, M.; Pan, D.; Zheng, H.; Wang, C.; Xu, F.; Zakaria, W.; and Dai, G. 2023c. DCAE: A dual conditional autoencoder framework for the reconstruction from EEG into image. *Biomedical Signal Processing and Control*, 81: 104440.

Zheng, X.; Chen, W.; Li, M.; Zhang, T.; You, Y.; and Jiang, Y. 2020. Decoding human brain activity with deep learning. *Biomedical Signal Processing and Control*, 56: 101730.