

DreamDiffusion: High-Quality EEG-to-Image Generation with Temporal Masked Signal Modeling and CLIP Alignment

Yunpeng Bai^{1,5} , Xintao Wang³, Yan-Pei Cao⁴, Yixiao Ge², Chun Yuan¹ ,
and Ying Shan²

¹ Tsinghua Shenzhen International Graduate School, China

² ARC Lab, Tencent PCG, ³ Kuaishou Technology, ⁴ VAST

⁵ The University of Texas at Austin

<https://github.com/bbaaii/DreamDiffusion>

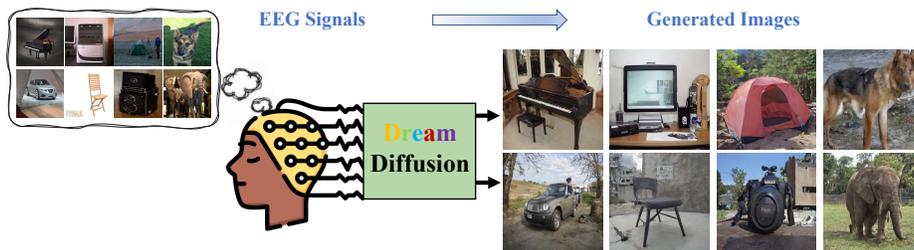


Fig. 1: Our proposed DreamDiffusion is capable of generating high-quality images directly from brain electroencephalogram (EEG) signals, without the need to translate thoughts into text.

Abstract. This paper introduces DreamDiffusion, a novel method for generating high-quality images directly from brain electroencephalogram (EEG) signals, without the need to translate thoughts into text. DreamDiffusion leverages pre-trained text-to-image models and employs temporal masked signal modeling to pre-train the EEG encoder for effective and robust EEG representations. Additionally, the method further leverages the CLIP image encoder to provide extra supervision to better align EEG, text, and image embeddings with limited EEG-image pairs. Overall, the proposed method overcomes the challenges of using EEG signals for image generation, such as noise, limited information, and individual differences, and achieves promising results. Quantitative and qualitative results demonstrate the effectiveness of the proposed method as a significant step towards portable and low-cost “thoughts-to-image”, with potential applications in neuroscience and computer vision.

Keywords: Image Generation · EEG Decoding · Model Pre-training

 Corresponding author.

1 Introduction

Image generation [4,15,21] has made great strides in recent years, especially after breakthroughs in text-to-image generation [1, 12, 29, 30, 33]. The recent text-to-image generation enables the creation of people’s ideas into exquisite paintings and artworks controlled by text. We are very curious whether we could control image creation directly from brain activities (such as electroencephalogram (EEG) recordings), without translating our thoughts into text before creation. This kind of “thoughts-to-images” has broad prospects and could broaden people’s imagination. For example, it can greatly improve the efficiency of artistic creation and help capture those fleeting inspirations. It also has the potential to help us visualize our dreams at night, (which inspires the name DreamDiffusion). Moreover, it may even aid in psychotherapy, having the potential to help children with autism and those with language disabilities.

Some recent works, such as MinD-Vis [7] and [39], attempt to reconstruct visual information based on fMRI (functional Magnetic Resonance Imaging) signals, which is another way to measure brain activities. They have demonstrated the feasibility of *reconstructing* high-quality results from brain activities. However, they are still far away from our goal of using brain signals to create conveniently and efficiently. 1) Since fMRI equipment is not portable and needs to be operated by professionals, it is difficult to capture fMRI signals. 2) The cost of fMRI acquisition is high. They greatly hinder the widespread use of this method in the practical artistic generation. In contrast, EEG (electroencephalogram) is a non-invasive and low-cost method of recording electrical activity in the brain. Portable commercial products are now available for the convenient acquisition of EEG signals, showing great potential for future art generation.

In this work, we aim to leverage the powerful generative capabilities of pre-trained text-to-image models (i.e., Stable Diffusion [31]) to generate high-quality images directly from brain EEG signals. However, this is non-trivial and has two challenges. **1)** EEG signals are captured non-invasively and thus are inherently noisy. In addition, EEG data are limited and individual differences cannot be ignored. *How to obtain effective and robust semantic representations from EEG signals with so many constraints?* **2)** Thanks to the use of CLIP [27] and the training on a large number of text-image pairs, the text and image spaces in Stable Diffusion are well aligned. However, the EEG signal has its own characteristics, and its space is quite different from that of text and image. *How to align EEG, text and image spaces with limited and noisy EEG-image pairs?*

To address the first challenge, we propose to train EEG representations using large amounts of EEG data instead of only rare EEG-image pairs. Specifically, we adopt masked signal modeling to predict the missing tokens based on contextual cues. Different from MAE [17] and MinD-Vis [7], which treat inputs as two-dimensional images and mask the *spatial information*, we consider the temporal characteristics of EEG signals, and dig deep into the semantics behind temporal changes in people’s brains. We randomly mask a proportion of tokens and then reconstruct those masked ones *in the time domain*. In this way, the pre-trained

encoder learns a deep understanding of EEG data across different people and various brain activities.

As for the second challenge, previous methods [7,39] usually directly fine-tune Stable Diffusion (SD) models using a small number of noisy data pairs. However, it is difficult to learn accurate alignment between brain signals (e.g., EEG and fMRI) and text spaces by end-to-end fine-tuning SD only using the final image reconstruction loss. We thus propose to employ additional CLIP [27] supervision to assist in the alignment of EEG, text, and image spaces. Specifically, SD itself uses *CLIP’s text encoder* to generate text embeddings, which are quite different from the masked pre-trained EEG embeddings in the previous stage. We leverage *CLIP’s image encoder* to extract rich image embeddings that align well with CLIP text embeddings. Those CLIP image embeddings are then used to further optimize EEG embedding representations. Therefore, the refined EEG feature embeddings can be well aligned with the CLIP image and text embeddings, and are more suitable for SD image generation, which in turn improves the quality of generated images.

Equipped with the above two delicate designs, our proposed method, namely, DreamDiffusion, can generate high-quality and realistic images from EEG signals. Our contributions can be summarized as follows. **1)** We propose DreamDiffusion, which leverages the powerful pre-trained text-to-image diffusion models to generate realistic images from EEG signals only. It is a further step towards portable and low-cost “thoughts-to-images”. **2)** We specifically explored temporal masked signal modeling method tailored to EEG data to learn effective representations, which is useful for subsequent related work. **3)** We further leverage the CLIP image encoder to provide extra supervision to better align the EEG, text, and image embeddings with limited EEG-image pairs. **4)** Quantitative and qualitative results have shown the effectiveness of our method.

2 Related works

2.1 Generating images from brain activity

The use of brain signals, including fMRI and EEG, to generate images has been an active area of research. For the use of fMRI, traditional methods rely on fMRI-image paired data to train the model to predict image features from fMRI. These image features will be fed into GANs [35] for stimulus reconstruction during testing. However, recent studies [3] have proposed unsupervised approaches, such as a reconfigurable autoencoder design, to learn from unpaired fMRI and images, and utilize regression models [24,26] to extract a latent fMRI representation that can be used to fine-tune a pre-trained conditional BigGAN [5] for decoding. The recent work MinD-Vis [8] integrates SC-MBM and DC-LDM to generate more plausible images with better-preserved semantic information.

Similarly, generating images from EEG signals has also been explored using deep learning techniques. Brain2image [22] have developed using LSTM and generative methods to learn a more compact representation of EEG data for generating visual stimuli that evoke specific brain responses. ThoughtViz [40] takes

encoded EEG signals as input to generate corresponding images, even with limited training data. [9] uses EEG as a supervision signal for learning semantic feature representations and achieving comparable performance to semantic image editing.

2.2 Model pre-training

Pre-training models have become increasingly popular in the field of computer vision, with various self-supervised learning approaches focusing on different pretext tasks [13, 25, 42]. These methods often utilize pretext tasks such as contrastive learning [2, 16], which models image similarity and dissimilarity, or autoencoding [6], which recovers the original data from a masked portion. In particular, masked signal modeling (MSM) has been successful in learning useful context knowledge for downstream tasks by recovering the original data from a high mask ratio for visual signals [17, 43] and a low mask ratio for natural languages [10, 28]. Another recent approach, CLIP [27], builds a multi-modal embedding space by pre-training on 400 million text-image pairs collected from various sources on the Internet. The learned representations by CLIP are extremely powerful, enabling state-of-the-art zero-shot image classification on multiple datasets, and providing a method to estimate the semantic similarity between text and images.

2.3 Diffusion models

Diffusion models have become increasingly popular as generative models for producing high-quality content [36]. The basic form of diffusion models is a probabilistic model defined by a bi-directional Markov Chain of states [18]. These models [11, 18, 32, 38] exhibit strong generative power due to their natural fit with the inductive biases of image-like data. The best synthesis quality is typically achieved when using a reweighted objective during training [18], allowing for a trade-off between image quality and compression capabilities. However, evaluating and optimizing these models in pixel space is computationally expensive and time-consuming [19, 23, 34, 37, 41]. To address these challenges, some diffusion models work on a compressed latent space of lower dimensionality, such as the proposed LDMs [31]. By compressing images into lower-dimensional latent features using a KL regularized autoencoder and then reconstructing them using the same latent space features, the LDM reduces computational costs while maintaining synthesis quality.

3 Proposed method

Our method comprises three main components: 1) masked signal pre-training for an effective and robust EEG encoder, 2) fine-tuning with limited EEG-image pairs with pre-trained Stable Diffusion, and 3) aligning the EEG, text, and image spaces using CLIP encoders. Firstly, we leverage masked signal modeling

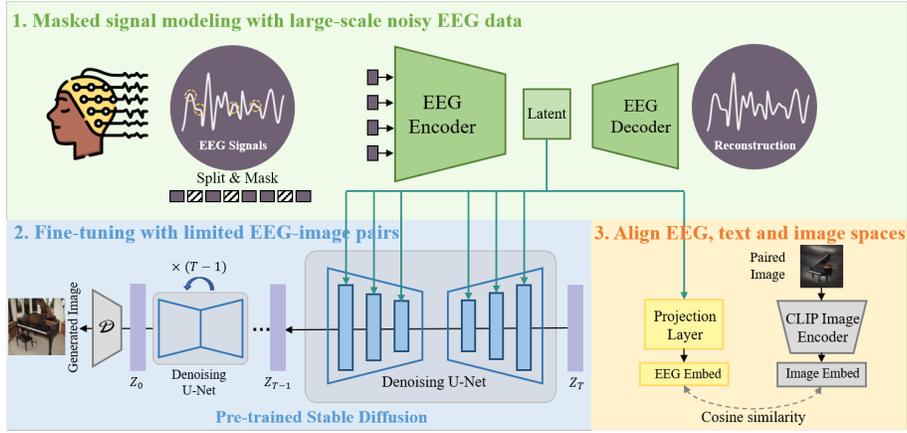


Fig. 2: Overview of DreamDiffusion. Our method comprises three main components: 1) masked signal pre-training for an effective and robust EEG encoder, 2) fine-tuning with limited EEG-image pairs with pre-trained Stable Diffusion, and 3) aligning the EEG, text, and image spaces using CLIP encoders.

with lots of noisy EEG data to train an EEG encoder to extract contextual knowledge. The resulting EEG encoder is then employed to provide conditional features for Stable Diffusion via the cross-attention mechanism. In order to enhance the compatibility of EEG features with Stable Diffusion, we further align the EEG, text, and image embedding spaces by reducing the distance between EEG embeddings and CLIP image embeddings during fine-tuning. After that, we obtain DreamDiffusion, which is capable of generating high-quality images from EEG signals only.

3.1 Masked signal pre-training for effective and robust EEG representations

EEG (Electroencephalogram) data is a recording of electrical activity generated by the human brain, measured using electrodes placed on the scalp. It is a non-invasive and low-cost method of measuring brain activity. EEG data has several characteristics. Firstly, the data is two-dimensional, with one dimension representing the channels or electrodes placed on the scalp, and the other dimension representing time. The temporal resolution of EEG is high, meaning that it can capture rapid changes in brain activity that occur on the order of milliseconds. However, the spatial resolution of EEG is low, meaning that it is difficult to precisely localize the source of the activity within the brain. Secondly, EEG signals are highly variable, influenced by factors such as age, sleep, and cognitive state. Finally, EEG data is often noisy, and requires careful processing and analysis to extract meaningful information.

Due to the inherent variability and noise in EEG data, conventional modeling methods often struggle to extract meaningful information from EEG signals.

The EEG data itself contains many signals unrelated to the corresponding visual concepts (noise), making the results of simply utilizing EEG through an unpretrained encoder inaccurate. Consequently, adopting masked signal modeling techniques, which have been proven effective in capturing contextual information from noisy and variable data [7, 17], represents a promising avenue for deriving meaningful contextual knowledge from large-scale noisy EEG data. Different from MAE [17] and MinD-Vis [7], which treat inputs as two-dimensional images and mask the *spatial information*, we consider the temporal characteristics of EEG signals, and dig deep into the semantics behind temporal changes in people’s brains.

Given the high temporal resolution of EEG signals, we first divide them into tokens in the time domain, and randomly mask a certain percentage of tokens. Subsequently, these tokens will be transformed into embeddings by using a one-dimensional convolutional layer. Then, we use an asymmetric architecture such as MAE [17] to predict the missing tokens based on contextual cues from the surrounding tokens. Through reconstructing the masked signals, the pre-trained EEG encoder learns a deep understanding of EEG data across different people and various brain activities.

3.2 Fine-tuning with Stable Diffusion on limited EEG-image pairs

After obtaining an effective representation of EEG signals from masked signal pre-training, we utilize it to generate images by leveraging a pre-trained Stable Diffusion (SD) model. Stable Diffusion involves gradually denoising a normally distributed variable to learn a data distribution. SD is augmented with a cross-attention mechanism for more flexible conditional image generation and the most common condition is the text prompt. Stable Diffusion has shown great generative power in generating high-quality images from various types of signals, such as labels, text, and semantic maps.

Stable Diffusion operates on the latent space. Given an image x in pixel space, x is encoded by a VQ encoder $\mathcal{E}(\cdot)$ to obtain the corresponding latent $z = \mathcal{E}(x)$. Conditional signals are introduced by the cross-attention mechanism in the UNet. This cross-attention can also incorporate conditional information from the EEG data. Specifically, the output of EEG encoder y is further projected with a projector τ_θ into an embedding $\tau_\theta(y) \in \mathbb{R}^{M \times d_\tau}$. Then, this EEG representation is incorporated into U-Net by a cross-attention layer implementing $\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V$.

$$Q = W_Q^{(i)} \cdot \varphi_i(z_t), K = W_K^{(i)} \cdot \tau_\theta(y), V = W_V^{(i)} \cdot \tau_\theta(y), \quad (1)$$

where $\varphi_i(z_t) \in \mathbb{R}^{N \times d_e^i}$ denotes intermediate values of the U-Net. $W_V^{(i)} \in \mathbb{R}^{d \times d_e^i}$, $W_Q^{(i)} \in \mathbb{R}^{d \times d_\tau}$ and $W_K^{(i)} \in \mathbb{R}^{d \times d_\tau}$ are projection matrices with learnable parameters.

During the fine-tuning process, we optimize the EEG encoder and cross-attention heads of the U-Net together. We keep the remaining parts of Stable Diffusion fixed. We use the following SD loss function for fine-tuning.

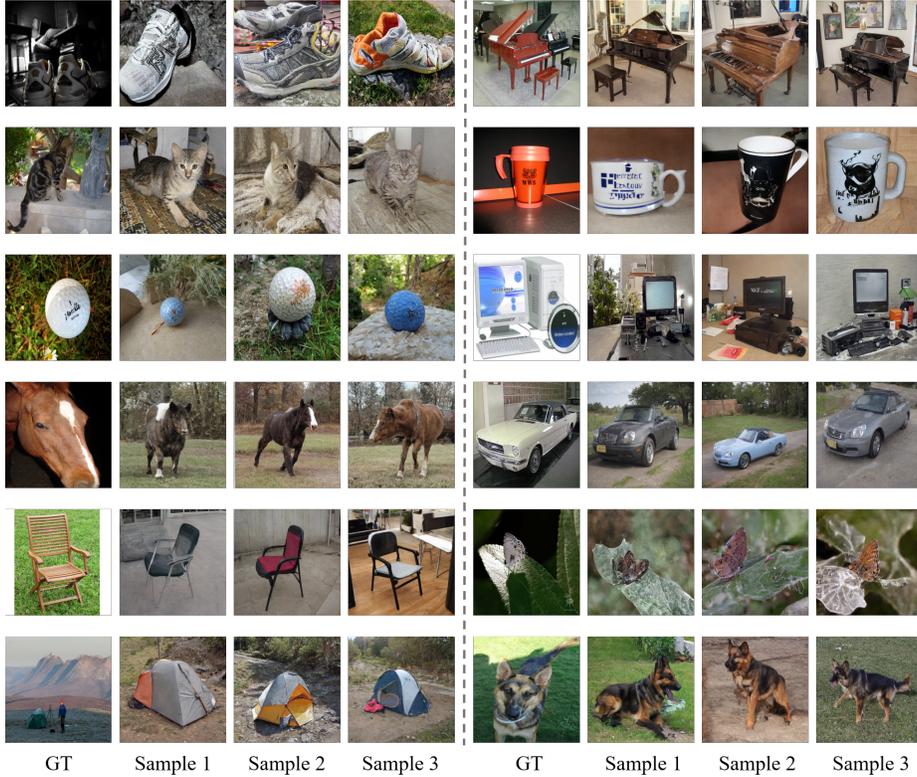


Fig. 3: Main results. The images on the left depict paired image data, while the three images on the right represent the sampling results. It can be observed that our model generates images of high quality from the EEG data, and these images match the EEG data accurately.

$$L_{SD} = \mathbb{E}_{x, \epsilon \sim \mathcal{N}(0,1), t} \left[\|\epsilon - \epsilon_{\theta}(x_t, t, \tau_{\theta}(y))\|_2^2 \right], \quad (2)$$

where ϵ_{θ} is the denoising function implemented as UNet.

3.3 Aligning the EEG, text, and image spaces with CLIP encoders

Next, we will fine-tune the EEG representation obtained from pre-training to make it more suitable for generating images. The pre-trained Stable Diffusion model is specifically trained for text-to-image generation; however, the EEG signal has its own characteristics, and its latent space is quite different from that of text and image. Therefore, directly fine-tuning the Stable Diffusion model using limited EEG-image paired data is unlikely to accurately align the EEG features with the text embeddings.



Fig. 4: Comparison with Brain2Image. The quality of the generated images produced by DreamDiffusion is significantly higher than those generated by Brain2Image.

Thanks to the use of CLIP [27] and the training on a large number of text-image pairs, the text and image spaces in Stable Diffusion are well aligned. Therefore, we propose to employ additional CLIP [27] supervision to assist in the alignment of EEG, text, and image space. Specifically, the EEG features obtained from the pre-trained encoder are transformed into embeddings with the same dimension as those of CLIP through a projection layer. We then use a loss function to minimize the distance between the EEG embeddings and the image embeddings obtained from the CLIP image encoder. The CLIP model is fixed during the fine-tuning process. The loss function is defined as follows:

$$\mathcal{L}_{clip} = 1 - \frac{E_I(I) \cdot h(\tau_\theta(y))}{|E_I(I)||h(\tau_\theta(y))|}, \quad (3)$$

where h is a projection layer and E_I is the CLIP image encoder. This loss function can encourage the EEG features to become more closely aligned with the image and thus more similar to text features. In this way, we can align the EEG signal, text and image in one unified space. The optimized EEG embedding representation is more suitable for SD image generation, which in turn improves the quality of generated images. CLIP alignment is not primarily to enhance understanding of EEG, but rather to improve the adaptability of the effective EEG representations obtained from pretraining to Stable Diffusion.

4 Experiments and Analyses

4.1 Implementation details

Data for EEG representation pre-training. We have collected approximately 120,000 EEG data samples from over 400 subjects with channel ranges from 30 to 128 on the MOABB [20] platform for the EEG pre-training. These data include all EEG data with more than 30 channels in the platform. MOABB is a software package designed to facilitate the development of brain-computer interface (BCI) algorithms by providing a collection of publicly available EEG datasets in a common format, along with a suite of state-of-the-art algorithms. This platform enables researchers to easily validate new algorithms using automated statistical analysis, eliminating the need for time-consuming and unreliable data preprocessing. These data contain a wide variety of EEG data, including tasks such as looking at an object, motor imagery, and watching videos. Our goal is to learn universal representations from diverse EEG data, without specific requirements on the types of EEG data.

Due to variations in the equipment used for data acquisition, the channel counts of these EEG data samples differ significantly. To facilitate pre-training, we have uniformly padded all the data that has fewer channels to 128 channels by filling missing channels with replicated values. During the pre-training process, every 4 adjacent time steps are grouped into a token and each token is transformed into a 1024-dimensional embedding through a projection layer for subsequent masked signal modeling. The loss function calculates the MSE between the reconstructed and original EEG signals. The reconstruction is performed on the entire set of 128 channels as a whole, rather than on a per-channel basis. The loss is only computed on masked patches. The reconstruction is performed on the entire set of 128 channels as a whole, rather than on a per-channel basis. The decoder is discarded after pretraining.

Our data processing method included the Beta (15-31 Hz) and Gamma (32-70 Hz) bands, as they convey information involved in visual perception, as well as potentially useful frequencies. EEG signals below 5Hz typically only appear during deep sleep, so it is unlikely that useful signals below 5Hz would be present in the data we used. Since the data used in pre-training comes from different regions and devices with varying line noises, we did not consider this for uniform processing. Moreover, based on experiments from previous works, the impact of line noise is not significant.

Paired EEG-image data. We adopt the ImageNet-EEG [22] dataset for our “thoughts-to-image” experiments, which is a collection of EEG recordings obtained from 6 subjects while they were shown 2000 images belonging to 40 different categories of objects from the ImageNet dataset. Each category consisted of 50 images, and each image was presented for 0.5 seconds, followed by a 10-second pause for every 50 images. The EEG data were recorded using a 128-channel Brainvision EEG system, resulting in a total of 12000 128-channel EEG sequences. The dataset includes images of various objects, such as animals (dogs, cats, elephants, etc.), vehicles (airliners, bikes, cars, etc.), and everyday

objects (computers, chairs, mugs, etc.). More details can be found in the related reference [22].

Other details. We use version 1.5 of Stable Diffusion for image generation. The mask ratio for EEG signals is set to 75%. All EEG signals are filtered within the frequency range of 5-95 Hz. Subsequently, the signals are truncated to a common length of 512. The encoder is pre-trained for 500 epochs and finetuned with Stable Diffusion for another 300. The pre-training model for EEG is similar to ViT-Large [14]. The training and testing were conducted on the same subject, and all results presented in the paper were generated using data from Subject 4. For more results, please refer to the supplementary material.

Table 1: More comparison evaluation. In the context of comparing various quantitative metrics, our method outperforms previous approaches significantly across the board.

Methods	FID ↓	IS ↑	PSNR ↑	SSIM ↑	LPIPS ↓
Brain2Image	18.76	5.06	12.8	0.213	0.701
Ours	3.61	28.54	14.6	0.267	0.644

4.2 Comparison with Brain2Image

In this section, we present a comparison of our proposed approach with Brain2Image [22]. However, it presents results for only a few categories and does not provide a reference implementation. In light of this, we conducted a qualitative comparison of the results on a few categories (namely, Airliner, Jack-o-Lantern, and Panda) that were showcased in the Brain2Image paper. To ensure a fair comparison, we followed the same subjective evaluation strategy as outlined by Brain2Image and presented generated instances of different methods in Figure 4. The top rows depict the results generated by Brain2Image, whereas the bottom rows were generated by our proposed method, DreamDiffusion. We observed that the quality of the generated images produced by DreamDiffusion is significantly higher than those generated by Brain2Image, thus validating the efficacy of our proposed method.

Due to the limited number of images available, using metrics like FID/IS may be unstable and cannot effectively measure the quality of generated images. Therefore, we provide the FID/IS metrics here only as a reference (Table 1). The metrics are calculated using images from Brain2Image paper. Nonetheless, our method significantly surpasses previous approaches in terms of these quantitative metrics. We also add some other similarity-based metrics or perception-based ones here. Since our method aims not for precise image reconstruction but for generating corresponding conceptual images from EEG signals, these metrics are provided for reference only.

4.3 Ablation studies

In this section, we conduct several ablation studies on the proposed framework using various cases. We evaluate the effectiveness of different methods by employing a 50-way top-1 accuracy classification task. We use a pre-trained ImageNet1K classifier [14] to determine the semantic correctness of the generated images. Both the ground-truth and generated images will be inputted into the classifier. Then, we will verify whether the top-1 classification of the generated image matches the ground-truth classification in 50 selected classes. A generated image will be deemed correct as long as the semantic classification results of the generated image and the ground-truth are consistent.

Table 2: Quantitative results of ablation studies. E and A represent fine-tuning of the encoder and cross-attention heads, respectively. Params: trainable parameters in the EEG encoder. Lines 1-4 mainly aim to illustrate the impact of clip fine-tuning without pre-training on the results. Lines 5-7 aim to explain the influence of different mask ratios on the results. Lines 8-11 demonstrate the impact of encoder parameter volume, where larger parameters do not necessarily yield better results. Finally, lines 12-14 elucidate the effects of fine-tuning the encoder or attention layer separately.

Model	MSM Pretraining	CLIP Finetuning	Mask Ratio	E + A	Params	Acc (%)
Full	✓	✓	0.75	E + A	297M	45.8
1	✗	✗	-	E + A	297M	4.2
2	✗	✗	-	E + A	18.3M	3.7
3	✗	✓	-	E + A	297M	32.3
4	✗	✓	-	E + A	18.3M	24.5
5	✓	✓	0.25	E + A	297M	19.7
6	✓	✓	0.5	E + A	297M	38.3
7	✓	✓	0.85	E + A	297M	33.4
8	✓	✓	0.75	E + A	458M	38.5
9	✓	✓	0.75	E + A	162M	36.6
10	✓	✓	0.75	E + A	74M	29.8
11	✓	✓	0.75	E + A	18.3M	28.7
12	✓	✓	0.75	E only	297M	22.4
13	✓	✗	0.75	E + A	297M	28.3
14	✓	✗	0.75	A only	297M	20.9

Role of pre-training: To demonstrate the effectiveness of the pretraining with large-scale EEG data, we conduct a validation by training several models with untrained encoders. One of the models is identical to the full model, while the other model has a shallow EEG encoding layer with only two layers to avoid overfitting the data. During the training process, the two models were trained with and without clip supervision, and the results are shown in Table 2, Model 1-4. It can be observed that the accuracy of the model without pre-training decreased.

The EEG data itself contains many signals unrelated to the corresponding visual concepts (noise), making the results of simply utilizing EEG through an unpretrained encoder inaccurate. The pretraining method, with random masking operations, exposes the model to various types of noise and interference during pre-training to reconstruct the original signal, thereby enabling the model to learn effective representations of visual concepts. The improvements in accuracy after using the pretrained encoder are manifested as “effective” and “robust.”

Mask ratios: We investigate to determine the optimal mask ratio for MSM pretraining with EEG data. As shown in Model 5-7 of Table 2, excessively high or low mask ratios can have a detrimental effect on the model’s performance. The highest overall accuracy was achieved at a mask ratio of 0.75. This finding is significant as it suggests that, unlike natural language processing where low mask ratios are commonly used, a high mask ratio is also a preferable option when performing MSM on EEG.

CLIP aligning: One of the keys of our method is to align the EEG representation with the image through the CLIP encoder. CLIP alignment is not primarily to enhance understanding of EEG, but rather to improve the adaptability of the effective EEG representations obtained from pretraining to Stable Diffusion. To validate the effectiveness of this approach, we conducted experiments 13-14 as shown in Table 2. It can be observed that the performance of the model significantly decreases when CLIP supervision is not used. In fact, as shown in the bottom right corner of Figure 5, even in the absence of pre-training, using CLIP to align EEG features can still yield reasonable results, which highlights the importance of CLIP supervision in our method.

Other aspects: We further illustrate the roles of the remaining parts by explaining Figure 5 and Table 2. As shown in the two images in the top-right corner of Figure 5, using an encoder without pre-training and fine-tuning with CLIP results in very poor quality generation. Even when the encoder is trained but not fine-tuned afterward, we find that it can decode concepts to some extent, but the quality is significantly inferior (bottom 1st). Without fine-tuning the cross-attention layers of the stable diffusion simultaneously, accurate results cannot be obtained (bottom 2nd). Decoding corresponding concepts without using CLIP during fine-tuning yields somewhat accurate but not entirely precise results (bottom 3rd). While fine-tuning solely with CLIP can align EEG with corresponding concepts to some extent, the effectiveness is not as high as with the fully pre-trained method (bottom 4th). In conclusion, achieving high-quality image generation from EEG data using Stable Diffusion is not straightforward, and each aspect of the method discussed in the paper is indispensable.

Looking at the results in Table 2, lines 1-4 primarily aim to illustrate the impact of clip fine-tuning without pre-training on the results. It can be observed that regardless of the number of parameters in the encoder, clip supervision helps construct the mapping from EEG signals to images, but it does not necessarily aid in learning effective EEG representations. In any case, the accuracy is lower compared to the pre-trained complete method. Lines 8-11 demonstrate the impact of encoder parameter volume, indicating that larger parameters do

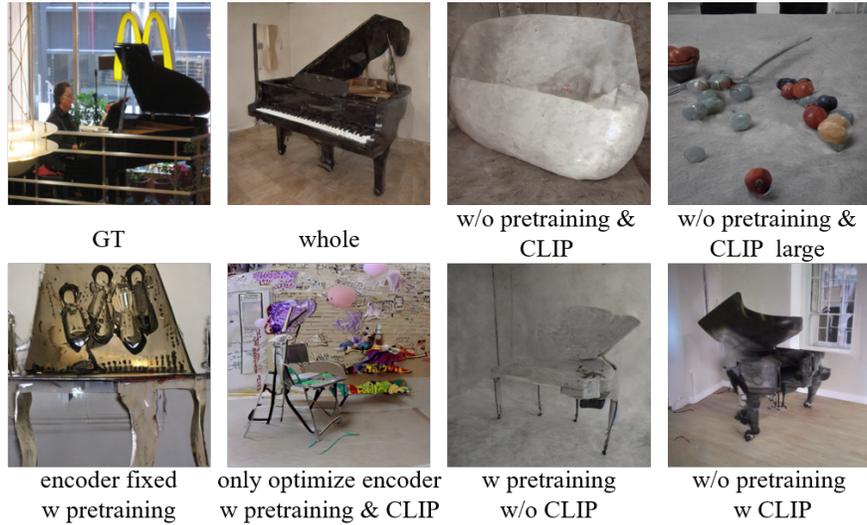


Fig. 5: Qualitative results of ablation studies. The top-right two images illustrate the effects of using an encoder without pre-training and fine-tuning with CLIP. The results when the encoder is not further fine-tuned after training (1st in the line below). The results when the cross-attention layers of stable diffusion are not fine-tuned simultaneously (2nd in the line below). The results when clip is not used during the fine-tuning process (3rd in the line below), and the results when only clip is used for fine-tuning (4th in the line below).

not always lead to better results. When the parameter volume is too large, it may fit some irrelevant noise in the EEG signals. Lines 12-14 aim to explain the effects of fine-tuning only the encoder or attention layer. It can be observed that fine-tuning either of them alone is not conducive to adapting the representation of EEG to the conditional input of stable diffusion, resulting in a decrease in corresponding accuracy.

4.4 Moving beyond coarse category information.

Comparing to reconstruction, our approach aims to utilize EEG as a conditional input to generate images. Sometimes, our brains conceive abstract concepts that are not specific entities. Like Brain2Image, our objective is to use EEG signals containing abstract concepts to generate corresponding images, rather than precise reconstruction for entities. Additionally, the data utilized in the current experiments mainly contain category-level information, as each image was displayed for 0.5 seconds during data acquisition. We will explore more detailed levels of image generation in the future, such as collecting EEG data with richer semantics.

Although EEG data only provide coarse-grained information at the category level in experimental results currently, our method aims to explore the possi-



Fig. 6: Failure cases of DreamDiffusion. Some certain categories are erroneously mapped to others due to their similarity in shapes or colors.

bility of generating images from EEG, rather than merely substituting category information with EEG. If we use category labels instead of CLIP as supervision, the accuracy will reach 86.7%. If we directly use category labels as input, the accuracy can reach 97.2%. However, adding category labels is not a good practice because we aim to use EEG in future that is semantically richer than just category information. Utilizing CLIP with EEG-image paired data for finetuning is undoubtedly a better and more appropriate choice for future application.

5 Conclusion

This paper proposes a novel method, DreamDiffusion, for generating high-quality images from EEG signals, which is a non-invasive and easily obtainable source of brain activity. The proposed method addresses the challenges associated with EEG-based image generation by utilizing the knowledge learned from large EEG datasets and the powerful generative capabilities of image diffusion models. Through a pre-training and fine-tuning scheme, EEG data can be encoded to the representation suitable for image generation using Stable Diffusion. Our method represents a significant advancement in the field of image generation from brain activity.

Limitations. Figure 6 shows some failure cases, where some categories are mapped to other categories with similar shapes or colors. We assume this may be due to the fact that the human brain considers shape and color as two important factors when recognizing objects. Nevertheless, DreamDiffusion has the potential to be used in a wide range of applications, such as neuroscience, psychology, and human-computer interaction.

Acknowledgements

This work was supported by the National Key R&D Program of China (2022YFB4701400/4701402), SSTIC Grant(KJZD20230923115106012, KJZD20230923114916032), and Beijing Key Lab of Networked Multimedia.

References

1. Bai, Y., Wang, C., Xie, S., Dong, C., Yuan, C., Wang, Z.: Textir: A simple framework for text-based editable image restoration. arXiv preprint arXiv:2302.14736 (2023)
2. Becker, S., Hinton, G.E.: Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature* **355**(6356), 161–163 (1992)
3. Bird, C.M., Berens, S.C., Horner, A.J., Franklin, A.: Categorical encoding of color in the brain. *Proceedings of the National Academy of Sciences* **111**(12), 4590–4595 (2014)
4. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096 (2018)
5. Brock, A., Donahue, J., Simonyan, K.: Large scale GAN training for high fidelity natural image synthesis. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net (2019)
6. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. *Advances in neural information processing systems* **33**, 1877–1901 (2020)
7. Chen, Z., Qing, J., Xiang, T., Yue, W.L., Zhou, J.H.: Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. arXiv preprint arXiv:2211.06956 (2022)
8. Chen, Z., Qing, J., Xiang, T., Yue, W.L., Zhou, J.H.: Seeing beyond the brain: Masked modeling conditioned diffusion model for human vision decoding. In: arXiv (November 2022), <https://arxiv.org/abs/2211.06956>
9. Davis, K.M., de la Torre-Ortiz, C., Ruotsalo, T.: Brain-supervised image editing. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 18480–18489 (2022)
10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
11. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems* **34**, 8780–8794 (2021)
12. Ding, M., Zheng, W., Hong, W., Tang, J.: Cogview2: Faster and better text-to-image generation via hierarchical transformers. arXiv preprint arXiv:2204.14217 (2022)
13. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: *Proceedings of the IEEE international conference on computer vision*. pp. 1422–1430 (2015)
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
15. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. *Communications of the ACM* **63**(11), 139–144 (2020)
16. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06). vol. 2, pp. 1735–1742. IEEE (2006)
17. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 16000–16009 (2022)

18. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* **33**, 6840–6851 (2020)
19. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.* **23**, 47:1–47:33 (2022)
20. Jayaram, V., Barachant, A.: Moabb: trustworthy algorithm benchmarking for bcis. *Journal of neural engineering* **15**(6), 066011 (2018)
21. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 4401–4410 (2019)
22. Kavassidis, I., Palazzo, S., Spampinato, C., Giordano, D., Shah, M.: *Brain2Image*: Converting brain signals into images. In: *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*. pp. 1809–1817. ACM (2017)
23. Kong, Z., Ping, W.: On fast sampling of diffusion probabilistic models. *CoRR* **abs/2106.00132** (2021), <https://arxiv.org/abs/2106.00132>
24. Mozafari, M., Reddy, L., VanRullen, R.: Reconstructing natural scenes from fmri patterns using bigbigan. In: *2020 International joint conference on neural networks (IJCNN)*. pp. 1–8. IEEE (2020)
25. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*. pp. 69–84. Springer (2016)
26. Ozcelik, F., Choksi, B., Mozafari, M., Reddy, L., VanRullen, R.: Reconstruction of perceived images from fmri patterns and semantic brain exploration using instance-conditioned gans. In: *2022 International Joint Conference on Neural Networks (IJCNN)*. pp. 1–8. IEEE (2022)
27. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *International conference on machine learning*. pp. 8748–8763. PMLR (2021)
28. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. *OpenAI blog* **1**(8), 9 (2019)
29. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022)
30. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: *International Conference on Machine Learning*. pp. 8821–8831. PMLR (2021)
31. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 10684–10695 (2022)
32. Ronneberger, O., Fischer, P., Brox, T.: 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. pp. 234–241. Springer (2015)
33. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., et al.: Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487* (2022)

34. San-Roman, R., Nachmani, E., Wolf, L.: Noise estimation for generative diffusion models. CoRR **abs/2104.02600** (2021)
35. Shen, G., Dwivedi, K., Majima, K., Horikawa, T., Kamitani, Y.: End-to-end deep image reconstruction from human brain activity. *Frontiers Comput. Neurosci.* **13**, 21 (2019)
36. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning. pp. 2256–2265. PMLR (2015)
37. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021)
38. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021)
39. Takagi, Y., Nishimoto, S.: High-resolution image reconstruction with latent diffusion models from human brain activity. *bioRxiv* pp. 2022–11 (2022)
40. Tirupattur, P., Rawat, Y.S., Spampinato, C., Shah, M.: Thoughtviz: Visualizing human thoughts using generative adversarial network. In: 2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018. pp. 950–958. ACM (2018)
41. Vahdat, A., Kreis, K., Kautz, J.: Score-based generative modeling in latent space. In: Ranzato, M., Beygelzimer, A., Dauphin, Y.N., Liang, P., Vaughan, J.W. (eds.) *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual.* pp. 11287–11302 (2021)
42. Wang, X., Gupta, A.: Unsupervised learning of visual representations using videos. In: *Proceedings of the IEEE international conference on computer vision.* pp. 2794–2802 (2015)
43. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simsim: A simple framework for masked image modeling. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* pp. 9653–9663 (2022)