Category-aware EEG Image Generation Based on Wavelet Transform and Contrast Semantic Loss

Enshang Zhang[1,*] , **Zhicheng Zhang**[2,3,*] , **Takashi Hanakawa**[1]
[1]Department of Integrated Neuroanatomy and Neuroimaging, Kyoto University Graduate School of Medicine, Kyoto, Japan.
[2]JancsiLab, JancsiTech, Hongkong, China
[3]Sino-Finland Joint AI Laboratory for Child Health of Zhejiang Province, Hangzhou, China
zhang.enshang.58p@st.kyoto-u.ac.jp, zhangzhicheng13@mails.ucas.edu.cn

Abstract

Reconstructing visual stimuli from EEG signals is a crucial step in realizing brain-computer interfaces. In this paper, we propose a transformer-based EEG signal encoder integrating the Discrete Wavelet Transform (DWT) and the gating mechanism. Guided by the feature alignment and category-aware fusion losses, this encoder is used to extract features related to visual stimuli from EEG signals. Subsequently, with the aid of a pre-trained diffusion model, these features are reconstructed into visual stimuli. To verify the effectiveness of the model, we conducted EEG-to-image generation and classification tasks using the THINGS-EEG dataset. To address the limitations of quantitative analysis at the semantic level, we combined WordNet-based classification and semantic similarity metrics to propose a novel semantic-based score, emphasizing the ability of our model to transfer neural activities into visual representations. Experimental results show that our model significantly improves semantic alignment and classification accuracy, which achieves a maximum single-subject accuracy of 43%, outperforming other state-of-the-art methods. The source code and supplementary material is available at https://github.com/zes0v0inn/DWT_EEG_Reconstruction/

1 Introduction

In recent years, the reconstruction of visual stimuli from electroencephalogram (EEG) has emerged as a highly promising research area within the domain of brain-computer interface (BCI), by extracting visually relevant features from EEG and ultimately reconstructing visual stimuli [Bai *et al.*, 2023; Kavasidis *et al.*, 2017]. This technology has the ability to convert neural signals into images, thereby establishing a crucial link between brain activities and the external world and deepening our comprehension of the intricate relationship between brain activities and perception. It holds immense potential, especially for individuals with severe disabilities. By enabling them to convey their thoughts and intentions through visual representations, it has the potential to revolutionize assistive communication methods.

With the powerful image-generation capabilities of generative networks, such as adversarial generative networks (GAN) [Goodfellow *et al.*, 2014], variational auto-encoder networks (VAE) [Kingma, 2013], and denoising diffusion probabilistic models (DDPM) [Zhang *et al.*, 2025; Ho *et al.*, 2020], which make it possible to perceive brain visual stimuli from EEG and reconstruct visual stimuli. However, existing related methods face substantial challenges. First, reconstructing visual stimuli from EEG at the pixel level is difficult and unnecessary. To this end, reconstructing semantically consistent images is of significant importance in BCI, leading to the inability to utilize traditional objective image evaluation indicators, such as structural similarity (SSIM). Therefore, the need arises for quantitatively evaluating the quality of visual stimuli to be reconstructed using an objective semantic-based score. In addition, EEG are time-series and noisy, which complicates the extraction of visually relevant features from them. Moreover, the training of advanced models, such as DreamDiffusion [Bai *et al.*, 2023] and MinD-vis [Chen *et al.*, 2023], requires an excessive amount of computational resources, severely restricting their widespread application.

To process noisy time-series EEG signals efficiently, the integration of traditional signal analysis methods, such as the discrete wavelet transform (DWT) [Chen *et al.*, 2017], into deep-learning modules has been employed [Zeng *et al.*, 2024]. By incorporating DWT-based modules into EEG encoder, we can leverage both the spatial and frequency characteristics of EEG, enhancing the feature-extraction process. In addition, with the continuous evolution of deep-learning models, the gate mechanisms in Mamba [Gu and Dao, 2023] have been proven effective in neural networks for selectively controlling information flow.

Motivated by the success of gated attention mechanisms in Mamba and DWT in signal processing, in this work, we integrated a DWT module with the gated attention mechanism within a well-designed EEG encoder, extracting meaningful features from EEG by effectively capturing both spatial and frequency-domain information while selectively focusing on relevant features. To reconstruct semantically consistent images without pixel-level ground truth, a category-aware clustering loss is utilized to cluster samples of the same category

[*] Corresponding Author

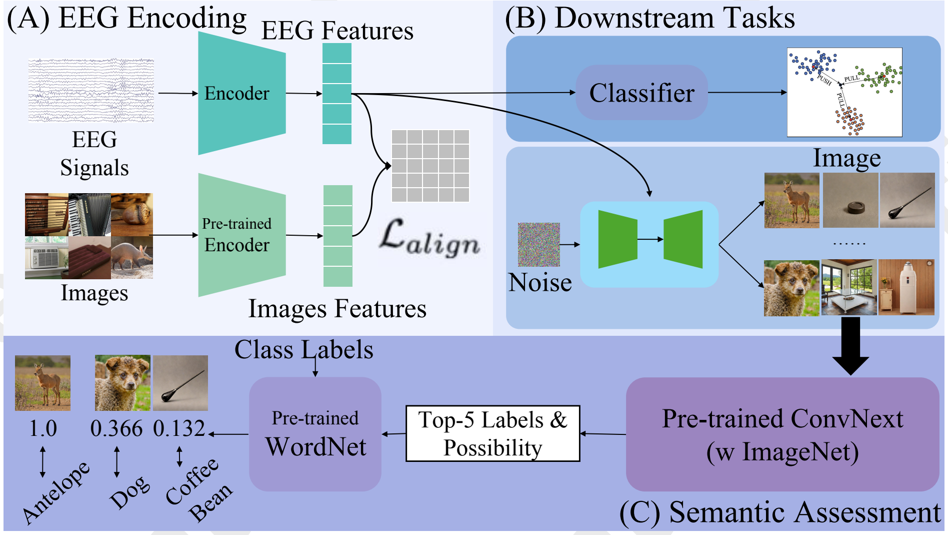

Figure 1: The overall flowchart of the proposed model. (A) EEG embedding part. (B) Downstream part. (C) Semantic evaluation part.

and separate different categories in a high-dimensional space, thereby improving the reconstruction accuracy and zero-shot generalization ability of the model. After the EEG feature extraction, we employ a pre-trained diffusion-based model to generate high-quality images. To ensure that the reconstructed images accurately reflect the semantic meaning of the input EEG, we incorporate an image classification model as a performance validation step. Subsequently, we propose a related evaluation metric of semantic-based score, enabling a quantitative assessment of the semantic consistency between the input EEG and the reconstructed images. The novelty of this work is three-fold:

1. To the best of our knowledge, it is the first time to use category-aware clustering loss to better extract the EEG features associated with visual stimuli, with the assistance of the CLIP loss [Radford *et al.*, 2021] to align image and EEG features.
2. We propose a novel semantic-based evaluation metric using pre-trained classifiers to quantitatively assess EEG-to-image reconstruction performance.
3. To address EEG's noisy time-series nature, we design a specialized encoder combining DWT and gated attention, achieving strong classification performance and supporting downstream image generation.

2 Related Work

Reconstructing visual stimuli from brain signals, such as functional magnetic resonance imaging(fMRI) and EEG, has achieved remarkable results in previous studies. For instance, conditional Generative Adversarial Networks and VAEs have been applied to encode fMRI signals and further reconstruct the visual stimuli from the features of brain signals. Studies such as Shen *et al.* [Shen *et al.*, 2019] and Takagi *et al.* [Takagi and Nishimoto, 2023] demonstrated that fMRI data could be translated into semantically correct and high-quality images. In addition, EEG-based reconstruction algorithms, despite the challenges of noise and lower spatial resolution, have also exhibited promising results. Techniques integrating convolutional neural networks with GANs, as in [Song *et al.*, 2021; Yang *et al.*, 2021], have enabled the generation of visual representations corresponding to real-time brain activity. To simplify the application of BCI and make the technology more low-cost and convenient for use, EEG is a more practical signal entry than fMRI. Recent outstanding results, such as DreamDiffusion [Bai *et al.*, 2023] and ATM-S [Li *et al.*, 2024], have shown that, with the assistance of the powerful image generation ability of diffusion model, deep learning models can reconstruct visual stimuli from EEG, capturing some semantic information embedded in neural activity. Moreover, multi-modal approaches combining EEG and fMRI have enhanced image quality and robustness by leveraging complementary features from both modalities. These advancements emphasize the growing potential of brain-signal-based image generation for applications.

3 Methods

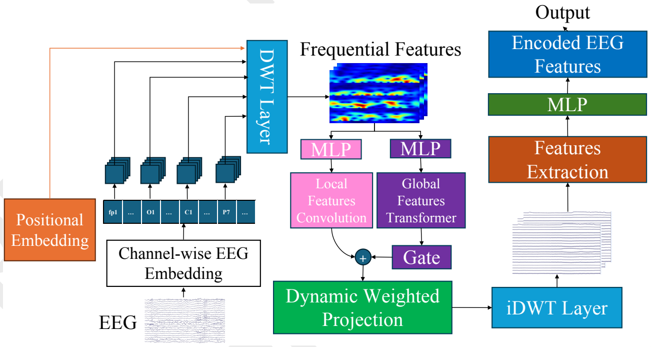

Figure 2: The overall structure of the encoder with several well-designed modules enhancing EEG embedding performance.

3.1 Overall Architecture

Fig. 1 depicts the overall flowchart of the proposed model. In this study, the proposed model comprises three main parts: EEG embedding with a well-designed EEG encoder seen in Fig. 2, the part of the downstream task (image reconstruction and EEG classification), and the semantic evaluation. Within the EEG embedding part, several well-designed modules are employed, including the DWT block, gated attention mechanisms, and a feature-fusion module. During network training, to achieve optimal performance in terms of feature alignment and category guidance, three distinct losses are combined: the CLIP loss, the Mean Square Error (MSE) loss, and the category-aware clustering loss. In the image reconstruction of the downstream task branch, the approach adopted is similar to that of the ATM-S method [Li *et al.*, 2024].

3.2 DWT Block

To effectively capture both the temporal- and frequency- domain characteristics of EEG, we propose a DWT-based feature extraction module utilizing the PyTorch Wavelet package [Cotter, 2020]. In this block, the EEG embedding input can be written as $E \in R^{B\times C\times T}$, where B represents the batch size, C denotes the number of channels and T is the length of the signal. Using DWT module, the model can synthesize both frequency and spatial features at different stages, enabling the model to learn more complex features.

Channel-wise DWT Decomposition: Given the time-series characteristics and high noise levels inherent in EEG, in this study, the Daubechies-1 (db1) wavelet known as the Haar wavelet, with a single-level decomposition is used for the implementation of our DWT module. The rationale behind this choice lies in the optimal localization properties of the db1 wavelet within the time domain, demonstrating remarkable effectiveness in capturing abrupt signal changes [Ocak, 2009]. Furthermore, single-level decomposition strikes an appropriate balance between frequency resolution and computational efficiency, which can balance the accuracy and computational demands for EEG processing. First, we apply one-dimensional (1D) DWT decomposition to each EEG channel independently:

$$\begin{aligned} [cA_1^c, cD_1^c] &= \mathrm{DWT}_{1D}(E_{\{b,c,:\}}) \\ s.t.\quad \forall c &\in \{1,...,C\}, b \in \{1,...,B\}, \end{aligned} \tag{1}$$

where $E_{\{b,c,:\}}$ denotes the temporal sequence for batch b and channel c. $cA_1^c \in R^{T/2}$ denotes the approximation coefficients of the low-frequency components. which can capture the overall trends and low-frequency patterns of the EEG. Concurrently, $cD_1^c \in R^{T/2}$ represents the detail coefficients of the high-frequency components, which retain the rapid changes and high-frequency characteristics of the signals.

Inverse DWT Module: We employ a 1D inverse DWT module (iDWT) to map the proposed EEG features back to the temporal domain according to the following equation:

$$F_{reconstructed} = \mathrm{iDWT}_{1D}([cA_1^c, cD_1^c], \psi), \tag{2}$$

where ψ represents the wavelet reconstruction parameters. This reconstruction process preserves not only the learned frequency-domain patterns but also temporal coherence. To integrate both the original temporal features and the optimized ones, we propose a feature fusion module, expressed mathematically as:

$$F_{fused} = Conv(W_f[F; F_{reconstructed}] + b_f), \tag{3}$$

where W_f and b_f are learnable parameters. F represents the proposed features derived from $[cA_1^c, cD_1^c]$, and $Conv$ denotes the convolution operations.

3.3 Gated Attention Mechanism

To process EEG embeddings from DWT modules while maintaining channel characteristics, we propose a gate-enhanced multi-head attention mechanism. This approach integrates the traditional gating mechanism within the attention of the transformer for selective feature processing. Our implementation features an adaptive gating integration mechanism for EEG embeddings, bridging local convolutional and global transformer features.

Local Feature Integration: The local feature extraction process begins with channel-wise temporal convolution, which processes each EEG channel independently, thus preserving their temporal relationships and individual patterns. Subsequently, we applied a cross-channel spatial convolution, integrating information from neighboring electrodes to effectively capture the spatial dependencies within the EEG structure. The integration of cross-channel features is realized through a comprehensive spatial processing pipeline that operates on wavelet-transformed signals. By applying two-dimensional convolution operations across the channel dimension of the wavelet coefficients, we effectively capture the intricate spatial relationships among different EEG channels within the frequency domain. This spatial integration process concatenates the wavelet coefficients from all channels before applying a spatial convolution kernel, enabling the model to learn meaningful cross-channel patterns. The convolution output then goes through batch normalization and non-linear activation, yielding a refined feature representation that simultaneously maintains the channel-specific frequency characteristics from the wavelet transform and the spatial relationships between channels.

Global Frequency Domain Attention: For global feature processing, we utilize a transformer-based encoder that employs multi-head self-attention mechanisms to capture long-range dependencies throughout the entire sequence length while preserving the channel-specific information structure. The integrated spatiotemporal features are then processed by a specialized channel-wise attention mechanism module. By applying a dedicated attention operation to each channel's spatiotemporal components, this module learns to selectively emphasize the most relevant patterns while simultaneously maintaining the unique characteristics of individual EEG channels. The use of multi-head attention assigns adaptive weights to different components based on their importance, enabling the model to focus on discriminative spectral features essential for downstream tasks while suppressing less informative frequency content.

Dynamic Weighted Projection: The integration process constructs a dynamic weighted combination model, where the local convolution features and the global transformer features are balanced based on the learned channel-specific weights. This method allows the model to automatically adjust the contribution ratio of different feature types according to the characteristics of the input signal and the specific requirements of each channel. As intelligent selectors, gate mechanisms excel in EEG processing by learning to balance local temporal patterns with global dependencies. The combination of local patterns and global dependencies has proven particularly valuable in different applications, including EEG processing [Song *et al.*, 2022; Lai *et al.*, 2023], since the functional role of different channels, determined by their spatial location and measured brain activity, often requires different processing strategies. Therefore, the relationship between local patterns and global signal dependence can vary considerably depending on different regions of the brain and cognitive states [Song *et al.*, 2022; Du *et al.*, 2023]. The adaptive specialty of our proposed gate mechanism ensures that the model can handle these changes while preserving the integrity of the signal. In addition, the combination of gating mechanisms with multi-head attention mechanisms offers a robust framework for EEG processing, enabling the model to concentrate on the most relevant patterns while maintaining computational efficiency.

3.4 Multi-Branch Feature Extraction Module

Due to the proposed multi-branch architecture, we effectively extract and fuse temporal and spatial features from EEG via parallel processing streams, capturing both time-domain characteristics and inter-channel relationships present in EEG. The temporal branch focuses on extracting time-domain characteristics from EEG by employing a series of well-designed convolution operations based on ShallowConvNet [Schirrmeister *et al.*, 2017]. First, a temporal convolution is applied across the time dimension while processing each channel independently, thereby capturing local temporal patterns during the task. Subsequently, the temporal features undergo dimensionality reduction through average pooling, and then a point-wise convolution is performed, which adjusts the feature representation while maintaining temporal relationships.

The spatial branch is designed to capture inter-channel relationships and spatial patterns across the EEG electrode array. We implement this by initially performing a spatial convolution across the channel dimension, enabling the network to learn local spatial patterns between neighboring electrodes. This is followed by depth-wise separable convolutions, which efficiently expand the receptive field while maintaining computational efficiency.

The fusion mechanism employs an attention-based approach to adaptively combine temporal and spatial features. After concatenating features from both branches, we apply a channel attention mechanism that generates dynamic weights for each feature channel. The attention module consists of channel-wise average pooling followed by two convolution layers with a bottleneck structure, producing attention weights through a sigmoid activation. The final stage of our architecture integrates the attention-weighted features through a fusion branch. A point-wise convolution combines the weighted features, followed by batch normalization and nonlinear activation. The output then undergoes a final adaptive pooling operation to ensure consistent dimensionality.

3.5 Loss Function

In our approach, we introduce a dual-loss mechanism that integrates three complementary components: CLIP loss and MSE loss for general feature alignment, and a category-aware clustering loss for enhanced clustering. By employing label-free category-aware clustering loss, we endow the model with potential zero-shot discrimination capabilities, particularly relevant in cases where the label space may differ between the training and testing phases, as observed in our dataset where the label categories of the training and testing sets are inconsistent. During network training, these two types of losses are combined using different weights as hyperparameters. We calculate the CLIP losses, including the loss between EEG features and image features and the loss between EEG features and text features, to ensure that the EEG features can reflect relevant information. Subsequently, we calculate the category-aware clustering loss for the classification results. The CLIP loss, $\mathcal{L}_{align}$ and the MSE loss, $\mathcal{L}_{MSE}$, between image features, F_I, and EEG features, F_E, can be written as follow:

$$\mathcal{L}_{align}(F_I^k, F_E^k) = \frac{1}{2}(\mathcal{L}_{I\rightarrow E}(F_I^k, F_E^k) + \mathcal{L}_{E\rightarrow I}(F_I^k, F_E^k)), \tag{4}$$

where k is the index of training sample.

Category-aware Clustering Loss: A key innovation of our approach lies in implementing contrastive learning without relying on absolute class labels through class-aware clustering loss. We dynamically construct relationships within each batch by examining the relative similarities among samples. Instead of depending on predefined classes, we adaptively create positive and negative pairs based on the data in the current batch, enabling the model to learn feature representations through comparison. This similarity-based clustering operation is independent of the total number of classes in the dataset, making it particularly effective in scenarios where class distributions may vary between training and testing phases. The category-aware clustering loss can be formulated as follow:

$$\mathcal{L}_C^k = \sum_{\substack{i=1 \\ i\neq j}}^{J} max(0, Sim(Q_j^k, C_i) - M) - log(Sim(Q_j^k, C_j) + \epsilon). \tag{5}$$

In the formulas above, by element-wise multiplication, the Sim refers to the similarity between the feature, Q_j^k, which belongs to the k^{th} sample of the j^{th} class, and the center of class j, C_j. J is the amount of categories in the current batch. ϵ is used as a constant to make sure the numerical stability during calculating. And the constant M is used as the threshold. When the dissimilarity achieves M, the loss assumes that these two samples could be completely separated. The final loss function can be written as follow:

$$\mathcal{L}_{loss} = \sum_{k=1}^{K} \lambda_1 \mathcal{L}_{align}(F_I^k, F_E^k) + \lambda_2 \mathcal{L}_{MSE}(F_I^k, F_E^k) + \lambda_3 \mathcal{L}_C^k, \tag{6}$$

where K is the number of samples in the current batch. λ_1, λ_2 and λ_3 are the weighting parameters to balance these three loss functions.

3.6 Dataset

We conducted our experiments using the THINGS-EEG [Gifford *et al.*, 2022] dataset, which is a large-scale collection of EEG recordings designed to capture neural signals to visual stimuli. The THINGS-EEG dataset was developed as part of the THINGS initiative[1], aiming at exploring the neural representation of object concepts. It comprises EEG recorded from participants who were exposed to a diverse set of real-world object images. These images were meticulously selected from the broader THINGS dataset [Hebart *et al.*, 2019], which encompasses over 26,000 object concepts spanning a wide semantic and perceptual range. In this work, MNE python package was applied to conduct the preprocessing. [Gramfort *et al.*, 2013]. The continuous EEG data were segmented into trials from $200ms$ before stimulus onset to $800ms$ after stimulus onset. Baseline correction was implemented by subtracting the mean pre-stimulus interval for each trial and channel. Subsequently, the data were downsampled to $100Hz$. Trials containing target stimuli were excluded from our experiment. For our experiments, we employed the same training and testing datasets from the THINGS-EEG dataset as those used in previous research.

3.7 Evaluation

In our experiments, we primarily used the classification accuracy including Top-1 accuracy and Top-5 accuracy to evaluate the capability of EEG encoder. For the reconstruction of visual stimuli, traditional quantitative metrics, such as MSE and SSIM, are not appropriate. This is because we only need to reconstruct images with consistent semantics and do not require pixel-level reconstruction recovery. Therefore, to assess the semantic information of the generated images, we creatively propose a semantic-based score as an evaluation metric for the quality of the generated images, based on the image classification model pre-trained on ImageNet, which, to the best of our knowledge, is the first time to be proposed. Specifically, we utilize the pre-trained weights of the ConvNext model from Meta [Liu *et al.*, 2022] to classify the generated images. For any generated image, we adopt the following rules:

1. If the result of Top-1 from the ConvNext model contains the same label, or the THINGS-EEG labels are included in the ImageNet labels, the score of the image is set to 1.
2. If the results of Top-5 from the ConvNext model contain the same label, the score of the image is set as the sum of the probabilities of the classes that contain the same label.
3. If the results of Top-5 from the ConvNext model do not contain the exact same label, we will use WordNet [Miller, 1995] to obtain the label's classes.

Then we calculate the score using the following formula:

$$Score = \sum P_i * Sim(label, WordNet_{label}) \tag{7}$$

Here, P_i represents the probability of each class obtained from the pre-trained ConvNext model, and Sim is the Wu-Palmer Similarity [Wu and Palmer, 1994].

Here, we carefully selected several representative methods, as comparison methods, including ATM-S, ATM-E [Li *et al.*, 2024], NERV [Chen, 2024], NICE [Song *et al.*, 2023], EIT-ResNet [Zheng *et al.*, 2024; He *et al.*, 2016]. Also, in the following part, we compared the generated images in the same subject, which is the subject "sub-08" to assess whether the generated images are able to reflect the semantic information successfully.

3.8 Implementation Details

In this work, we well-trained the proposed model with a maximum of 40 epochs, a learning rate of 3×10^{-4}, and a batch size of 64. Early stopping was applied, and training was halted if the top-1 classification accuracy did not improve for 10 consecutive epochs. To complement the evaluation, we also recorded the accuracy of the top 5 classification to capture the five most probable class predictions, which were further applied to analyze semantic consistency. In the whole training process, we also applied pre-trained models in different parts to reduce the computation cost and increase the generality. First, we applied the public CLIP weights to generate image embedding from the testing and training dataset of THINGS-EEG. Then, we applied the pre-trained diffusion models including SDXL-Turbo to generate images [Podell *et al.*, 2023; Sauer *et al.*, 2025; Ye *et al.*, 2023]. Finally, we used the pre-trained ConvNext trained with ImageNet [Deng *et al.*, 2009] and wordNet [Miller, 1995] to calculate the semantic scores based on the classification results.

4 Experimental Results

4.1 Visual Inspection of Image Reconstruction

To evaluate the ability of our framework to reconstruct high-quality images from EEG, we conducted relevant image-generation tasks on the THINGS-EEG dataset using all relevant model reconstructions. Fig. 3 exhibits the representative reconstructed images from subject-08. The first column presents the visual stimuli corresponding to the EEG signals. In terms of the semantic quality of the generated images, significant differences exist among the results produced by different methods shown in the subsequent columns. The second column displays the images generated by our method. From Fig. 3, we can see that When faced with the visual stimulus of "antelope", both our method and ATM-S retain, to a certain extent, the key features related to "antelope". Although pixel-level precise reconstruction may not be achieved, the overall outlines and main structures are highly recognizable. For example, the shape of the antelope can be roughly outlined. However, the images reconstructed by other methods

[1] https://things-initiative.org/

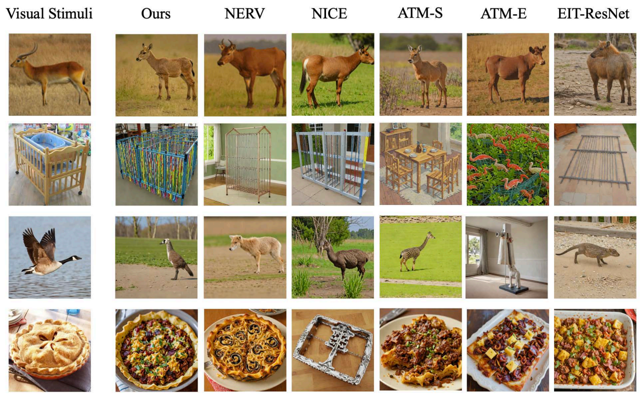

Figure 3: Examples from subject-08: left column shows EEG visual stimuli; subsequent columns display method reconstructions. Results demonstrate our model learns semantic features from EEG.

resemble either cows or horses. For another instance, considering the visual stimulus of ”bird” in the third row, only our method can, to some degree, make the generated image recognizable as a bird. Other methods fail to do so. In particular, the image reconstructed by ATM-E is not an animal-related image at all, showing a huge semantic discrepancy.

4.2 Quantitative Semantic Analysis

From Fig. 3, it can be observed that although our method cannot reconstruct visual stimuli at the pixel level, it is capable of generating images with similar semantics. To quantitatively compare the semantic similarity of images generated by different methods, in this work, we propose a quantitative semantic-based score. We categorize the scores of all generated images into three types based on the mean and standard deviation. For scores above (mean + std), the images are classified as good. For scores between (mean + std) and (mean - std), the images are classified as intermediate. For scores below (mean - std), the images are classified as bad. From Fig. 4, we can see that, to some extent, the higher the score, the closer the semantics of the reconstructed image is to the visual stimuli corresponding to the EEG signals. By comparing the magnitudes of these scores, we can, to a certain extent, conduct a horizontal comparison of the quality of images reconstructed by different methods, seen in Table 1.

Method	Mean	Standard deviation
Ours	**0.383**	0.182
NERV	0.376	0.198
NICE	0.364	0.167
ATM-S	0.368	0.176
ATM-E	0.326	0.154
EIT-ResNet	0.343	0.147

Table 1: Semantic scores of different models in subject-08.

4.3 Zero-shot EEG Classification

In the process of our EEG reconstruction of visual stimuli, a powerful EEG feature extractor is required to better extract corresponding semantic information from EEG. Evaluation of the ability of EEG feature extraction to extract relevant semantic information from EEG can be used to assess the ability of each model to reconstruct EEG visual stimuli. Therefore, we conducted a zero-shot classification task using the THINGS-EEG testing dataset. By aligning the EEG features with image features through the CLIP loss, we simultaneously optimize the EEG features via category-aware clustering loss, enabling the separation of different representations in the high-dimensional space while clustering similar ones. Subsequently, these features are processed by a lightweight Multi-Layer Perceptron (MLP) classifier.

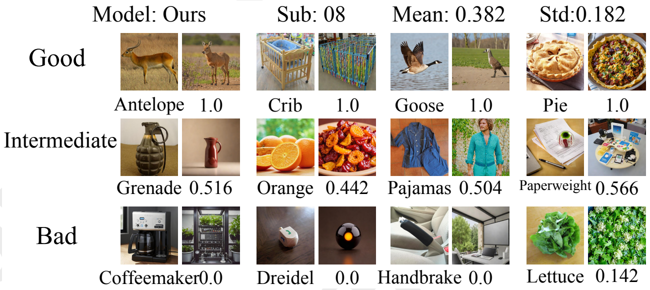

Figure 4: Figure shows subject-08’s semantic scores for generated images. Each pair: left = visual stimuli, right = generated result. Right image’s score is shown below. Classification: score > (mean + std) = good; score < (mean – std) = bad; others = intermediate.

The results, as presented in Table **??**, demonstrate that our model achieves superior performance compared to other baseline methods. To be specific, although our method does not achieve the best Top-1 and Top-5 accuracy across all subjects, overall, it demonstrates the best performance, which can be clearly observed from the average precision. This suggests that our method is more stable and effective in handling the task in general, rather than excelling in only a few specific cases.

4.4 Ablation Study

To explore the impact of each component on the proposed model, several ablation studies were conducted. First, based on our model, we removed the feature alignment loss. As shown in Table 2, $\mathcal{L}_{align}$ is the most crucial loss for training our model. Without $\mathcal{L}_{align}$, the accuracy of EEG classification becomes extremely low, indicating that the model cannot fit the data under such circumstances. Furthermore, we removed the DWT module, the local branch, and the global branch separately to train the classification task for each subject. The accuracy of the classification task decreased significantly across all subjects. A similar phenomenon also occurred in another ablation experiment: when we changed the gated attention from the global branch to the local branch. Regarding the other two loss functions $\mathcal{L}_{MSE}$ and $\mathcal{L}_{C}$ during the training process, we found that removing either one of them led to a slight decline in the performance of our model.

5 Discussion

EEG, as an indispensable tool in BCI, enables non-invasive acquisition of relevant brain information. EEG can be utilized to understand brain-related activities and guide clinical

Method	Sub-01		Sub-02		Sub-03		Sub-04		Sub-05		Sub-06		Sub-07		Sub-08		Sub-09		Sub-10		mean	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Ours	**33.0**	**58.5**	**28.0**	**56.5**	33.5	**61.0**	**36.0**	**68.0**	**26.0**	48.0	30.5	62.5	**34.0**	**62.5**	**43.0**	73.5	**31.5**	58.5	**38.5**	69.0	**33.4**	**61.8**
NERV	25.5	57	26	57	31.5	60.5	30.0	58.0	23.0	**50.0**	**32.0**	55.5	31.0	61.0	40.5	70.5	30.5	60.5	34.0	66.0	30.4	59.64
NICE	25.0	49.0	17.5	45.5	30.5	55.5	33.0	60.5	14.0	37.5	28.5	54.0	24.5	52.5	39.0	70.5	23.5	45.5	30.0	62.0	26.55	53.25
ATM-S	24.0	56.5	24.5	53.0	**34.5**	60.0	35.0	65.0	20.5	48.5	31.0	**65.5**	31.5	**62.5**	40.5	**75.0**	30.0	**59.0**	**38.5**	**70.0**	31.0	61.5
ATM-E	19.0	49.5	18.5	40.0	28.5	59.5	31.0	56.0	17.5	39	23.5	52.0	25.5	53.0	30.5	64.0	24.5	49.0	32.0	60.5	25.05	52.25
EIT-ResNet	16.5	30.5	11.5	26.5	13.0	38.0	14.0	32.0	8.0	24.5	15.5	39.0	16.5	40.0	16.0	39.5	12.5	26.0	14.5	43.5	13.8	34.0

Table 2: The overall accuracy for the classification task. We implemented the models from previous papers including NERV, ATM-S, ATM-E, *etc.* it can be seen that our proposed model still advance in both mean Top-1 and top-5 accuracy.

	$\mathcal{L}_{align}$	$\mathcal{L}_{MSE}$	$\mathcal{L}_{C}$	DWT	Local branch	Global branch	Sub-01	Sub-02	Sub-03	Sub-04	Sub-05	Sub-06	Sub-07	Sub-08	Sub-09	Sub-10	mean
Gated attention in global branch	✗	✓	✓	✓	✓	✓	1.0	0.5	1.0	0.5	1.0	0.5	1.0	1.0	0.5	1.0	0.8
	✓	✗	✓	✓	✓	✓	29.5	26.0	31.0	34.0	19.0	28.0	32.0	38.5	28.5	35.0	30.15
	✓	✓	✗	✓	✓	✓	30.0	27.0	31.5	37.5	22.5	31.5	34.5	42.5	31.5	39.5	32.8
	✓	✓	✓	✗	✓	✓	20.5	25.0	30.5	29.0	20.0	25.5	24.5	42.5	25.5	30.5	27.35
	✓	✓	✓	✓	✗	✓	23.5	20.5	28.5	24.0	17.0	26.0	24.5	34.5	23.0	26.0	24.75
	✓	✓	✓	✓	✓	✗	21.0	26.5	33.0	28.0	20.5	27.5	24.5	40.0	27.0	27.0	27.5
Gated attention in local branch	✓	✓	✓	✓	✓	✓	20.5	26.5	32.0	30.5	21.0	28.5	26.5	41.5	23.0	27.5	27.75
Ours							**33.0**	**28.0**	**33.5**	**36.0**	**26.0**	**30.5**	**34.0**	**43.0**	**31.5**	**38.5**	**33.4**

Table 3: The results for the ablation study in classification task.

treatment and assisted living for relevant individuals. Among them, the reconstruction of relevant visual stimuli from EEG can make EEG more tangible, which is conducive to connecting individuals with the real world.

In this work, we present a novel framework for EEG-to-image reconstruction, which enhances the extraction of meaningful features from EEG. To be specific, the proposed model integrates advanced components such as the DWT block and gated attention mechanisms, as well as a novel mixed loss function that combines CLIP loss for feature alignment with a label-free category-aware clustering loss, aiming to improve classification accuracy and semantic alignment. Furthermore, we innovatively utilize WordNet-based classification accuracy and semantic similarity measures and creatively propose a semantic-based score to objectively evaluate the semantic information of generated images, addressing limitations in existing evaluation methods and providing a scalable and quantitative way to assess the semantic consistency of reconstructed images.

Compared with previous results, EIT [Zheng *et al.*, 2024] applied mature deep learning tools such as ResNet as the EEG feature extractor for image reconstruction. However, despite its advance in efficiency and computation cost, the nature of such computer vision models may not be suitable for processing complex time-series data, especially for EEG data and the complicated downstream tasks such as high-level EEG feature extraction and feature alignment with images. Compared to NERV [Chen, 2024] and ATM-S [Li *et al.*, 2024], our model successfully introduced the DWT module as a frequency-domain feature extractor. Also, the application of category-aware clustering loss leads to further performance improvements.

Based on our analysis, there is a strong correlation between the classification accuracy and the semantic score of the generated images. Specifically, EEG that are correctly classified by the model typically result in images with higher semantic similarity scores. In contrast, EEG that are not correctly classified usually produce images with significantly lower semantic similarity scores. This finding emphasizes the importance of accurate classification in maintaining semantic fidelity during the EEG-to-image conversion process. Challenges found when generating images for specific categories: Some categories completely fail to generate semantically meaningful images. We have identified several potential reasons for this limitation: (1) Ambiguous or incorrect labels: Categories with unclear or incorrect labels, such as ”bator4”, pose significant challenges. Even when converted into semantic vectors through CLIP, these labels lack meaningful semantic relationships, leading to image generation failures. (2) Diffusion model constraints: In this paper, the directly used pre-trained diffusion model for reconstructing visual stimuli may lack corresponding visual representations for certain categories, limiting its ability to generate relevant images.

Despite numerous challenges, this study highlights the potential of EEG-based image generation as a research frontier. Future work could explore the use of interpretable models or traditional EEG analysis methods to extract more robust features, thereby enhancing the interpretability of the process. Moreover, improving the consistency between EEG features and semantic representations through more sophisticated architectures or loss functions could further advance the field. Developing customized pre-training strategies for diffusion models to include a broader range of categories is another promising avenue for improving image generation results.

In conclusion, our findings demonstrate the feasibility and value of EEG-based image generation, offering insights into the semantic encoding of neural signals and paving the way for more interpretable and accessible brain-computer interface technologies.

Ethical Statement

There are no ethical issues.

Acknowledgments

This work was partially supported by the Beijing Natural Science Foundation (L245015), the National Key R&D Program Program of China (2023YFC2706400), the Chongqing Natural Science Foundation (CSTB2024NSCQ-MSX0451), and the Medical Innovation Program(MIP) in Kyoto University. The authors also would like to show a great appreciation towards Dr.Yoshifumi Mori, who now works in Netherlands Institute for Neuroscience, for his valuable suggestions and insights in psychiatric diseases.

References

[Bai *et al.*, 2023] Yunpeng Bai, Xintao Wang, Yan-pei Cao, Yixiao Ge, Chun Yuan, and Ying Shan. Dreamdiffusion: Generating high-quality images from brain eeg signals. *arXiv preprint arXiv:2306.16934*, 2023.

[Chen *et al.*, 2017] Duo Chen, Suiren Wan, Jing Xiang, and Forrest Sheng Bao. A high-performance seizure detection algorithm based on discrete wavelet transform (dwt) and eeg. *PloS one*, 12(3):e0173138, 2017.

[Chen *et al.*, 2023] Zijiao Chen, Jiaxin Qing, Tiange Xiang, Wan Lin Yue, and Juan Helen Zhou. Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22710–22720, 2023.

[Chen, 2024] Chi-Sheng Chen. Necomimi: Neural-cognitive multimodal eeg-informed image generation with diffusion models. *arXiv preprint arXiv:2410.00712*, 2024.

[Cotter, 2020] Fergal Cotter. *Uses of complex wavelets in deep convolutional neural networks*. PhD thesis, 2020.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[Du *et al.*, 2023] Changde Du, Kaicheng Fu, Jinpeng Li, and Huiguang He. Decoding visual neural representations by multimodal learning of brain-visual-linguistic features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10760–10777, 2023.

[Gifford *et al.*, 2022] Alessandro T Gifford, Kshitij Dwivedi, Gemma Roig, and Radoslaw M Cichy. A large and rich eeg dataset for modeling human visual object recognition. *NeuroImage*, 264:119754, 2022.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[Gramfort *et al.*, 2013] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A. Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, and Matti S. Hämäläinen. MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, 7(267):1–13, 2013.

[Gu and Dao, 2023] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

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

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[Kavasidis *et al.*, 2017] Isaak Kavasidis, Simone Palazzo, Concetto Spampinato, Daniela Giordano, and Mubarak Shah. Brain2image: Converting brain signals into images. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1809–1817, 2017.

[Kingma, 2013] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[Lai *et al.*, 2023] Zhi-Hao Lai, Tian-Hao Zhang, Qi Liu, Xinyuan Qian, Li-Fang Wei, Song-Lu Chen, Feng Chen, and Xu-Cheng Yin. Interformer: Interactive local and global features fusion for automatic speech recognition. *arXiv preprint arXiv:2305.16342*, 2023.

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

[Liu *et al.*, 2022] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

[Miller, 1995] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[Ocak, 2009] Hasan Ocak. Automatic detection of epileptic seizures in eeg using discrete wavelet transform and approximate entropy. *Expert Systems with Applications*, 36(2):2027–2036, 2009.

[Podell *et al.*, 2023] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

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

[Sauer *et al.*, 2025] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pages 87–103. Springer, 2025.

[Schirrmeister *et al.*, 2017] Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human brain mapping*, 38(11):5391–5420, 2017.

[Shen *et al.*, 2019] Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. Deep image reconstruction from human brain activity. *PLoS computational biology*, 15(1):e1006633, 2019.

[Song *et al.*, 2021] Yonghao Song, Lie Yang, Xueyu Jia, and Longhan Xie. Common spatial generative adversarial networks based eeg data augmentation for cross-subject brain-computer interface. *arXiv preprint arXiv:2102.04456*, 2021.

[Song *et al.*, 2022] Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. Eeg conformer: Convolutional transformer for eeg decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2022.

[Song *et al.*, 2023] Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding natural images from eeg for object recognition. *arXiv preprint arXiv:2308.13234*, 2023.

[Takagi and Nishimoto, 2023] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14453–14463, 2023.

[Wu and Palmer, 1994] Zhibiao Wu and Martha Palmer. Verb semantics and lexical selection. *arXiv preprint cmp-lg/9406033*, 1994.

[Yang *et al.*, 2021] Delong Yang, Dongnan Su, Zhaohui Luo, Peng Shang, and Zhigang Hu. The survey of image generation from eeg signals based on deep learning. In *2021 International Symposium on Biomedical Engineering and Computational Biology*, pages 1–5, 2021.

[Ye *et al.*, 2023] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

[Zeng *et al.*, 2024] Yan Zeng, Jun Li, Zhe Zhao, Wei Liang, Penghui Zeng, Shaodong Shen, Kun Zhang, and Chong Shen. Wet-unet: Wavelet integrated efficient transformer networks for nasopharyngeal carcinoma tumor segmentation. *Science Progress*, 107(2):00368504241232537, 2024.

[Zhang *et al.*, 2025] Youjian Zhang, Li Li, Jie Wang, Xinquan Yang, Haotian Zhou, Jiahui He, Yaoqin Xie, Yuming Jiang, Wei Sun, Xinyuan Zhang, et al. Texture-preserving diffusion model for cbct-to-ct synthesis. *Medical Image Analysis*, 99:103362, 2025.

[Zheng *et al.*, 2024] Xu Zheng, Ling Wang, Kanghao Chen, Yuanhuiyi Lyu, Jiazhou Zhou, and Lin Wang. Eit-1m: One million eeg-image-text pairs for human visual-textual recognition and more. *arXiv preprint arXiv:2407.01884*, 2024.