Abstract

Reconstructing visual images from brain signals is a major challenge in brain-computer interfaces. Functional magnetic resonance imaging (fMRI) captures complex neural activity in the visual cortex. However, the extreme scarcity of fMRI data hinders learning of generalizable feature representations. The high-dimensional nature of fMRI signals combined with their low signal-to-noise ratio makes learning generalizable feature representations particularly challenging from limited samples. Furthermore, most existing models perform poorly in cross-subject scenarios due to significant individual variability in brain anatomy and function that cannot be easily standardized through conventional alignment methods. To address these issues, we propose BrainCanvas, a novel network designed to tackle both data scarcity and cross-subject variability. BrainCanvas leverages diffusion models and customized conditional prompts to augment training images, generating a diverse imageset with consistent semantics. Simultaneously, we design a Domain-Invariant Encoder with a Domain Adapter and adversarial training, allowing the model to adapt to cross-subject fMRI data. By learning subject-invariant feature representations shared among individuals, the model effectively reduces the impact of individual differences and enables cross-subject transfer. Experiments demonstrate that BrainCanvas achieves meaningful zero-shot cross-subject decoding performance, and with efficient fine-tuning on minimal target data, further attains highly competitive results in high-fidelity visual reconstruction, showcasing significant potential in the field of brain decoding.

Key words

fMRI decoding,Brain visual model,Cross-subject transfer

1 Introduction

In brain-computer interface (BCI) research, accurately reconstructing human visual perception non-invasively remains a major challenge [1]. Visual perception involves processing external stimuli and cognitive activities, generating complex neural signals in the visual cortex [2]. These signals, captured by techniques like fMRI, provide valuable insights into human cognition and decision-making [3]. Decoding such neural activities reveals how the brain processes visual stimuli, advancing neuroscience and cognitive science.

With the continuous development of fMRI, it has become possible to capture real-time brain activity signals with higher precision and finer temporal resolution [4]. By incorporating pre-trained visual models, fMRI decoding has become more effective in learning detailed feature embeddings from limited datasets [5, 6]. By combining pre-trained decoders such as latent diffusion models [7,8] and GANs [9], as well as pixel detail priors and CLIP [10] semantic condition priors, researchers can generate realistic and semantically consistent images.

However, existing methods face significant challenges. First, data scarcity is a critical issue, as fMRI data is costly and limited by experimental and ethical constraints [11]. Deep learning models often require large datasets, and insufficient data can lead to overfitting, reducing generalizability. Moreover, no suitable data augmentation methods for fMRI exist, as simple transformations can distort spatial and functional information due to fMRI's high dimensionality, low signal-to-noise ratio, and individual variability [12]. Secondly, existing decoding models are often subject-specific due to significant inter-individual differences in brain functional activity and anatomical structure. These differences make it difficult to capture shared features across individuals, limiting cross-subject generalization [13]. Thus, obtaining diverse training features and achieving cross-subject brain decoding are crucial challenges in cognitive neuroscience and BCI [14].

To address these issues, we propose BrainCanvas, which aims to enhance fMRI-based visual decoding by generating semantically diverse training imageset and achieving cross-subject feature alignment. Research from Neuron [15] showed the brain categorizes scenes based on statistical co-occurrence of objects in natural environments. Despite visual variations, the brain effectively categorizes this diverse visual input into distinct, representative semantic categories. Furthermore, research [16] found that stimuli perceived as similar elicit similar neural activation patterns. These studies suggest image variants retaining key semantic features—despite visual differences—evoke similar neural responses. BrainCanvas uses diffusion models to augment training imageset, preserving semantic consistency while introducing visual diversity to improve feature learning and generalization. We introduce a Category-aware Localization Module and Class-Prompt Appending to ensure generated images align semantically with the originals, overcoming data scarcity without distorting fMRI data.

For cross-subject decoding, we design a Domain Adapter to align features across subjects through domain adaptation representation learning [17] and adversarial training mechanisms, thus alleviating individual differences. The Domain Adapter uses a discriminator to distinguish individual features, with a generator producing features indistinguishable to the discriminator. This process allows the model to capture consistent cross-subject features, improving generalization. Consequently, when facing new subjects, the model demonstrates strong zero-shot cross-subject transfer capability and can achieve great performance with minimal fine-tuning, significantly enhancing practical feasibility. This also addresses the limitation of existing cross-subject methods that optimize only between pairs of subjects, lacking global consistency constraints, and improves performance in cross-subject tasks.

In summary, our contributions are as follows:

- **Proposing BrainCanvas to overcome data scarcity**: BrainCanvas generates semantically consistent yet visually diverse imageset, enhancing generalization on limited data while avoiding fMRI data distortion. BrainCanvas represents the first work to address both data scarcity and inter-subject variability in one framework. Our contribution extends beyond simply proposing another model; we present a general method that can promote



Fig. 1 fMRI-based brain decoding and image reconstruction. Left: Overview of visual decoding tasks under natural image stimulus. BrainCanvas augments the training dataset using Semantic Diversification ImageSet Generation. Our model processes fMRI signals collected after subjects view images. The Domain-Invariant Encoder (DIE) learns shared invariant features across multiple subjects. Using these features, the model reconstructs images, enabling cross-subject learning and transfer. Right: Our reconstruction examples.

the development of the fMRI community.

- **Enhancing cross-subject adaptability**: The Domain Adapter enables learning of subject-invariant features through domain-adaptive representation learning with adversarial training. This approach alleviates transfer difficulties due to individual differences and improves performance in cross-subject tasks.
- **Demonstrating superior experimental results**: Extensive experiments verify the effectiveness of BrainCanvas, showcasing significant advantages in high-fidelity, cross-subject human visual reconstruction, proving the innovation and practical value of our method.

2 Related Work

2.1 Neural signal decoding and reconstruction

Traditional image decoding methods extract features from fMRI data using pre-trained convolutional neural networks (CNNs), which are then fed into generative models for reconstruction [18, 19]. Horikawa and Kamitani [20] proposed a preprocessing method based on machine vision principles, successfully extracting visual features. Architectures like BigBiGAN [21] and IC-GAN [22] optimized fMRI latent space representations, though generated images still suffer from distortion. Recent methods combine high-resolution image synthesis with latent diffusion models [7, 23] and CLIP [24, 25], mapping fMRI signals to the CLIP embedding space for more precise reconstruction. Wang et al. [26] innovated cross-subject brain decoding, but their cyclic loss optimizes only pairwise subject data, lacking global consistency. However, these methods require large paired stimulus-response data, and individual differences pose challenges. Accurately decoding and reconstructing novel images remains a pressing problem.

2.2 Dataset Generation

In recent years, dataset generation methods have become popular in visual tasks, enabling the creation of large annotated datasets to address extensive annotation requirements, especially in tasks like semantic segmentation [27, 28] and image classification [29]. He et al. [30] demonstrated the effectiveness of synthetic data in zero-shot and few-shot image classification. Google's research team introduced the Imagen text-to-image generation model [31], achieving SOTA performance by fine-tuning on class-conditional ImageNet [32]. Nguyen et al. [33] proposed a diffusion model-based approach for generating pixel-level semantic segmentation labels with stable diffusion, reducing the need for labor-intensive pixel annotations. However, in the field of brain signal decoding and reconstruction, data augmentation methods using diffusion models have not been explored.

2.3 Multi-Subject Visual Decoding

The substantial data requirements for individual subjects represent a critical bottleneck in single-subject visual decoding models, driving researchers to increasingly pursue cross-subject approaches. Recent advances in multi-subject brain visual decoding have concentrated on two key challenges: aligning fMRI data across different individuals and capturing inter-subject variations. Various alignment strategies have emerged to address these challenges. For instance, MindBridge [26] employs adaptive fMRI aggregation techniques to harmonize neural responses across subjects. Building upon similar objectives, MindEye2 [34] takes a different approach by utilizing simple subject-specific ridge regressions for cross-subject alignment. Meanwhile, UMBRAE [35] tackles a distinct but related problem by attempting to decode grounding information from visual content within fMRI signals. Furthermore, Wills Aligner [36] leverages the fsaverage template as a standardized framework for cross-subject data alignment. More recently, BrainGuard [37] introduces a privacy-aware perspective, specifically engineered to improve image reconstruction from multi-subject fMRI data while simultaneously protecting individual privacy. Although these methods achieve better decoding performance than the single-subject methods, they predominantly emphasize inter-subject differences while neglecting to systematically learn inter-subject commonalities that could further enhance generalization capabilities.

3 Method

In this section, we present a detailed description of the proposed BrainCanvas architecture, as illustrated in Figure 3. The architecture consists of two modules: Imageset Diffusion and Multimodal Embedding Extractor, and fMRI-to-Image Reconstruction. We will delve into these two modules to reveal the working mechanisms and innovations of

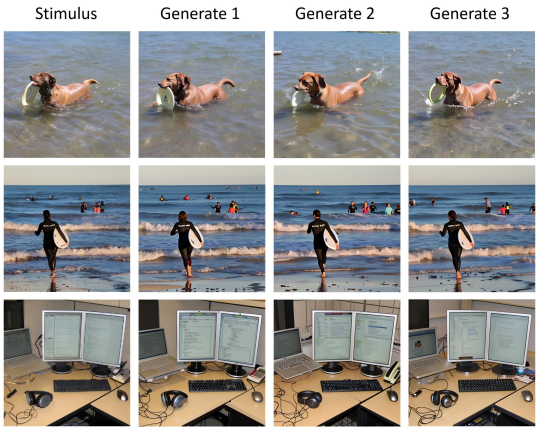


Fig. 2 Diverse images with similar semantics generated for the same brain signal. This illustrates the concept of Imageset Diffusion Samples.

BrainCanvas in the task of visual reconstruction.

3.1 Imageset Diffusion and Multimodal Embedding Extractor

To tackle the issue of data scarcity in brain decoding tasks, a direct approach is to employ data augmentation techniques. However, the complexity and nonlinearity of fMRI signals make traditional methods unsuitable, as they primarily perturb low-level pixel statistics but fail to model semantic consistency for fMRI-to-image mapping, while direct transformations risk distorting spatial-functional information. Despite these challenges, brain signals can exhibit stable semantic features under certain conditions. By leveraging these features, we can generate diverse images with similar semantics for the same brain signal, as illustrated in Figure 2. The generated images share core semantics but vary in details. This forces the model to focus on fMRI features correlated with high-level semantics rather than overfitting to low-level pixel patterns. Building on the existing NSD dataset, we describe training images using those paired with identical neural signals and associated descriptive information. Our goal is to enable the model to reconstruct images that closely match the input visual stimuli and accurately restore the original images' salient features.

To achieve this, we adopt a **Class-prompt appending** technique to fully utilize the category labels provided in the dataset. Specifically, for each image i, we concatenate its caption S_i with its class label C_i, which lists all object categories in the image as $C_i = \{c_1, \ldots, c_M\}$, where M is the number of categories. For example, for the input image in Figure 3, the combined text prompt L_i would be: *"Two men with surfboards walk towards the ocean; Two men, surfboards, ocean"*.

To ensure the generated imageset maintains structural consistency with the originals, we introduce a **Category-aware Localization Module**. This module's primary function is to compute a Perceptual Position Map (P_i) that provides explicit spatio-semantic guidance. Consequently, the diffusion process is conditioned on three synergistic sources of information: the class-appended text prompt L_i, the original image I_i and this computed guidance map P_i.

First, Class-prompt appending text prompt L_i is embedded into a text vector $e_L \in \mathbb{R}^{\Lambda \times d_e}$ via a text encoder, where Λ is the text length and d_e is the embedding dimension. Simultaneously, the original image I_i is transformed into a latent feature representation $z \in \mathbb{R}^{H \times W \times d_z}$ with an image encoder, with H and W representing image dimensions and d_z is the number of latent feature channels.

At each denoising step t, the latent features z_t are transformed into denoised features z_{t-1} using a UNet-based architecture that performs self-attention and cross-attention computations at layer l. The self-attention map $A_S^{l,t}$ captures similarities between different pixels in z_t^l, enhancing global contextual information. The cross-attention map $A_C^{l,t}$ models the relationship between the text embeddings e_L and the latent image features z_t^l, aligning text prompts with image regions to better understand the correspondence between category information and visual features. Formally, $A_S^{l,t}$ and $A_C^{l,t}$ are computed as:

$$A_S^{l,t} = \text{Softmax}\left(\frac{Q_z K_z^\top}{\sqrt{d_l}}\right) \quad A_C^{l,t} = \text{Softmax}\left(\frac{Q_z K_e^\top}{\sqrt{d_l}}\right) \tag{1}$$

where Q_z and K_z are the query and key obtained from the latent features z, and K_e is the key from text embeddings e_L, respectively, obtained by linear projections and taken as inputs to the attention mechanisms, and d_l is the feature dimension at layer l. After computing the attention maps at each time step t and layer l, we average them over layers and time steps to obtain global attention representations A_S and A_C:

$$A_S = \frac{1}{L \times T} \sum_{l=1}^{L} \sum_{t=0}^{T} A_S^{l,t}, \quad A_C = \frac{1}{L \times T} \sum_{l=1}^{L} \sum_{t=0}^{T} A_C^{l,t} \tag{2}$$

where L is the total number of layers and T is the total number of denoising steps.

However, a simple combination of these global maps is often insufficient to balance structural and semantic details effectively. This necessitates a more advanced strategy.

To form a more robust Perceptual Position Map P_i, we therefore employ a Hierarchical Attention-based Fusion mechanism. In this paradigm, A_C acts as a dynamic spatial filter to refine A_S, amplifying features in object-relevant regions while suppressing them elsewhere. This process is formulated as:

$$\begin{aligned} A_S^{\text{refined}} &= A_S + \tanh(\text{Conv}\alpha(A_S)) \odot \sigma(\text{Conv}\beta(A_C)) \\ P_i &= \text{Conv}_{\text{final}}([A_S^{\text{refined}}; A_C]) \end{aligned} \tag{3}$$

The refinement leverages a residual connection, where the original map A_S is updated by a semantically gated signal. The refined structural map A_S^{refined} is then concatenated with A_C and processed by a final convolutional layer $\text{Conv}_{\text{final}}$ to produce the hierarchically fused guidance map P_i.

Finally, a Structure Adapter injects this guidance map P_i into the intermediate layers of the diffusion model. This provides direct spatial conditioning throughout the generation process. Using the combined guidance from the text prompt L_i and the map P_i, we perform diffusion on the original imageset to generate a diverse set of new images. These augmented images and their prompts are then processed by a pre-trained CLIP model to extract target embeddings for the subsequent fMRI reconstruction task. This strategy ensures the augmented data is both visually and semantically consistent with the original, providing a richer dataset to support brain decoding.

3.2 fMRI-to-Image Reconstruction

This section addresses brain signal processing for reconstructing visual images from fMRI data. To effectively address cross-subject brain

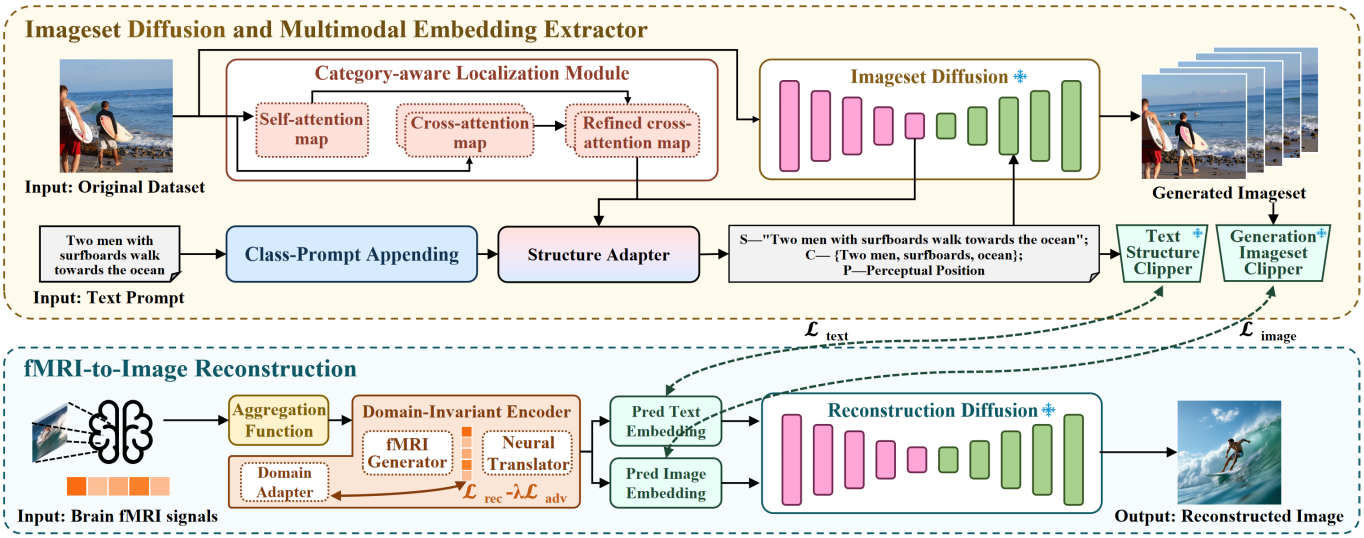


Fig. 3 Schematic Diagram of BrainCanvas. Upper Section: Augments the original imageset using the Category-aware Localization Module and Class-Prompt Appending, then extracts corresponding image and text embeddings. Lower Section: Processes fMRI signals from different subjects with the Domain-Invariant Encoder to obtain shared feature representations. These embeddings are combined with a diffusion model to generate the reconstructed images.

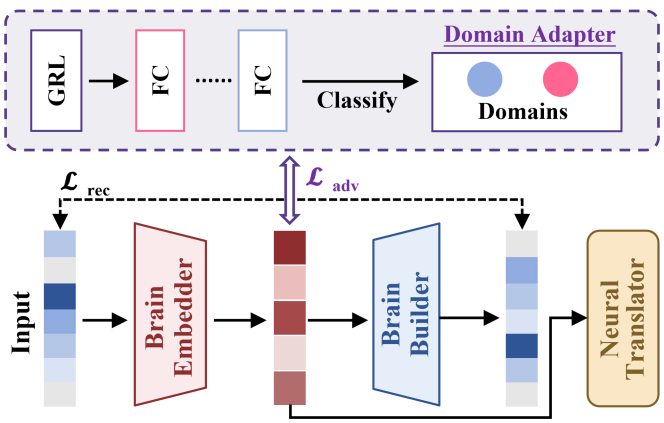


Fig. 4 The Domain-Invariant Encoder (DIE). The encoder learns subject-invariant features by combining a self-supervised reconstruction loss with adversarial training against a Domain Adapter. This removes subject-specific variability from the fMRI embeddings.

decoding, we propose a Domain-Invariant Encoder structure based on the brain decoding framework [26]. Existing methods often struggle to learn shared features across individuals. Moreover, current cyclic loss functions in cross-subject approaches tend to be local and lack global consistency, limiting their ability to capture subject-invariant features. To overcome this, we introduce an adversarial mechanism that encourages features extracted from different subjects to have similar distributions—effectively learning subject-invariant features. These shared representations enable transfer learning by mitigating individual differences.

Specifically, the model employs an aggregation function f_{agg} to adaptively map input fMRI voxels to an 8192-dimensional latent space: $h = f_{\text{agg}}(X)$. X represents the input fMRI signals and h is the aggregated feature representation. The fMRI Generator module then processes h through the Brain Embedder f_{emb} to generate a 2048-dimensional subject-invariant semantic embedding e_s:

$$e_s = f_{\text{emb}}(h; \theta_{\text{emb}}) \in \mathbb{R}^{d_s}, \quad d_s = 2048, \tag{4}$$

where θ_{emb} denotes the learnable parameters of the Brain Embedder.

To optimize e_s for capturing consistent features across individuals, we employ two strategies:

Self-Supervised Reconstruction Task: The embedding e_s is processed by a Brain Builder module, f_{recon}, to reconstruct the original normalized brain signals. The reconstruction loss, L_{rec}, ensures that e_s preserves sufficient signal-specific information by minimizing the L2 distance between the original and reconstructed signals:

$$L_{\text{rec}} = \mathbb{E}_{(X,s)\sim\mathcal{D}}\left[\frac{1}{N_s}\left\|X - f_{\text{recon}}(e_s)\right\|_2^2\right], \tag{5}$$

where $\mathcal{D}$ denotes the training dataset and N_s is the number of voxels for subject s. This self-supervised learning reinforces signal alignment across subjects, aiding the model in learning consistent features and enhancing cross-subject performance.

Adversarial Domain Adaptation: To enhance the subject invariance of e_s, we design an adversarial training process using a Domain Adapter, aiming to reduce bias in the distribution of features due to individual differences. Specifically, e_s is processed by through a Gradient Reversal Layer (GRL) [38] and then through fully connected layers f_{adv}, yielding the discriminator's output $\hat{d}$:

$$\hat{e}_s = \text{GRL}_\eta(e_s), \quad \hat{d} = f_{\text{adv}}(\hat{e}_s; \theta_{\text{adv}}), \tag{6}$$

where η is the GRL coefficient and $\hat{d} \in \mathbb{R}^K$ predicts the probability that e_s belongs to each of the K subjects. The Domain Adapter aim to minimizes the domain classification loss $\mathcal{L}_{\text{adv}}$:

$$\mathcal{L}_{\text{adv}} = -\mathbb{E}_{(e_s,d)\sim\mathcal{D}}\left[\sum_{k=1}^{K} d_k \log \hat{d}_k\right], \tag{7}$$

with d_i being the true subject labels. During training, the feature extractor f_{emb} and the Domain Adapter f_{adv} are optimized adversarially. The optimization objective for the feature extractor is to min-max:

$$\min_{\theta_{\text{emb}}} \max_{\theta_{\text{adv}}} (\mathcal{L}_{\text{rec}} - \lambda \mathcal{L}_{\text{adv}}) \tag{8}$$

Due to the effect of the GRL, the gradient of $\mathcal{L}_{\text{adv}}$ is changed when updating f_{emb}, which drives the feature extractor to maximize the domain classification loss $\mathcal{L}_{\text{adv}}$. This mechanism hinders the discriminator from correctly distinguishing data from different individuals, thereby enforcing the feature extractor to learn subject-invariant features. The domain adapter aims to accurately classify subject labels of the features. Through this adversarial training, the feature extractor learns shared representations across subjects, reducing individual differences , which enhances model performance in cross-subject applications.

After optimizing e_s, we use Neural Translator f_{trans} to generate two target embeddings: predicted CLIP image embedding $\hat{e}_I$ and text embedding $\hat{e}_T$. These are compared with the true embeddings e_I and e_T to optimize the model. We incorporate the SoftCLIP loss [25], which measures alignment in CLIP space and effectively aligns fMRI modalities to the pre-trained CLIP embedding space. The mean squared error loss $\mathcal{L}_{\text{MSE}}$ quantifies the difference between predicted and true embeddings:

$$\mathcal{L}_{\text{image}} = \mathcal{L}_{\text{SoftCLIP}}(\hat{e}_I, e_I) + \mathcal{L}_{\text{MSE}}(\hat{e}_I, e_I) \tag{9}$$

$$\mathcal{L}_{\text{text}} = \mathcal{L}_{\text{SoftCLIP}}(\hat{e}_T, e_T) + \mathcal{L}_{\text{MSE}}(\hat{e}_T, e_T) \tag{10}$$

The overall loss $\mathcal{L}$ balances terms with hyperparameters λ, α, β:

$$\mathcal{L} = \mathcal{L}_{\text{rec}} - \lambda \mathcal{L}_{\text{adv}} + \alpha \mathcal{L}_{\text{image}} + \beta \mathcal{L}_{\text{text}}, \tag{11}$$

The optimized embeddings $\hat{e}_I$ and $\hat{e}_T$ are inputs to a Stable Diffusion model [39], reconstructing visual images akin to the originals. Through this end-to-end process, BrainCanvas achieves cross-subject brain decoding by learning subject-invariant features, significantly enhancing visual reconstruction quality across subjects. This method provides a novel solution for decoding brain signals across individuals, effectively reducing individual variation effects on model performance.

4 Experiments

4.1 Dataset and Evaluation Metrics

Dataset: In this study, we employed the Natural Scenes Dataset (NSD) [40], a publicly available fMRI dataset featuring high-resolution 7-Tesla measurements from eight healthy adults who passively viewed thousands of MS-COCO [41] natural scenes across 30-40 scans. Following common practices [6, 24–26, 42–44], our research uses data from 4 subjects(1,2,5,7), who completed all the scan sessions. A common subset of 982 images viewed by all four was designated as the test set, while the remaining 8,859 images viewed individually by each participant were used for training. Additionally, we augmented the NSD dataset using semantic diversification imageset diffusion techniques, with detailed results presented in the ablation studies.

Evaluation Metrics: To enable quantitative comparisons with existing methods, we employ eight image quality assessment metrics.

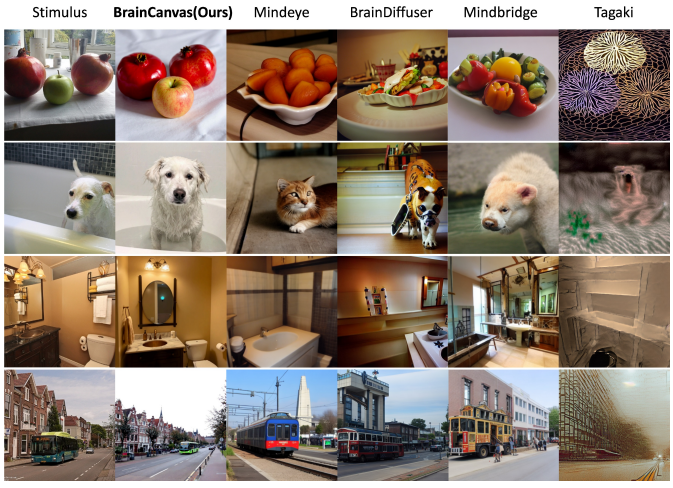


Fig. 5 Visualization results and comparison with other methods on the NSD dataset.

For low-level evaluation, PixCorr and SSIM [45] assess pixel-wise and structural similarities, while AlexNet (layer 2) and AlexNet (layer 5) [46] evaluate feature similarity across hierarchical levels. For high-level evaluation, Inception [47] and CLIP [10] measure semantic consistency, with CLIP focusing on multimodal alignment. Additionally, EfficientNet-B [48] and SwAV [49] use self-supervised learning to extract high-level features and assess image alignment. These metrics ensure a comprehensive evaluation from pixel fidelity to semantic coherence.

4.2 Implementation Details

In the Imageset Diffusion phase, we processed the original imageset using the PyTorch framework and the Stable Diffusion 2.1-base model [7] on an NVIDIA RTX 4090 GPU. Specifically, we set the time step $T = 100$ to control the temporal scale of the diffusion process, balancing computational efficiency and generation quality. The parameter L corresponds to the number of layers in the UNet network within the Stable Diffusion model. For cross-attention and self-attention maps, we set their feature scales to 16 and 32, respectively, to strike a balance between capturing detailed information and preserving global structure. To generate the final refined cross-attention map, we set the parameter $\tau = 4$. During optimization, we employed the AdamW optimizer with a learning rate of 1×10^{-4} and a weight decay coefficient of 1×10^{-4}.

In the fMRI-to-Image Reconstruction phase, the image reconstruction component relies on the Versatile Diffusion model [39]. We performed 20 diffusion steps to achieve an optimal balance between computational cost and reconstruction quality. The guidance scale was set to 5 to control the influence of conditioning information on the generation process. The text-to-image ratio was set to 0.5. In designing the loss function, we set $\lambda = 0.1$, $\alpha = 1$, and $\beta = 1 \times 10^4$. Similarly, we employed the AdamW optimizer with a learning rate of 1.5×10^{-4} and a weight decay coefficient of 1×10^{-2}. For the BrainCanvas model used in cross-subject experiments, we trained for 800 epochs comprising multiple subjects. For the new subject adaptation experiments, we fine-tuned the existing model for 200 epochs using data from the new

subject. This process enables the model to rapidly adapt to individualized neural patterns, enhancing the decoding accuracy of brain activity for new subjects.

4.3 Comparison with Previous Works

To evaluate BrainCanvas's performance, we trained several comparative networks, including Takagi *et al.* [24], Mind-Reader [50], MindVis [51], Gu *et al.* [42], Brain-Diffuser [6], MindEye [25], MindEye2 [34], MindBridge [26], Wills Aligner [36] and UMBRAE [35]. Additionally, we trained a Single Subject model using data from one participant without cross-subject pre-training or transfer learning, and averaged the results. In the table, BrainCanvas augmented the original dataset by generating twice data at a 512×512 resolution through ImageSet Diffusion. As shown in Table 1, BrainCanvas demonstrates superior performance across multiple key metrics. In cross-subject evaluations, BrainCanvas outperforms MindBridge on all metrics, particularly excelling in Inception and CLIP scores, even surpassing all single-subject models. This indicates that BrainCanvas-generated images are more consistent with visual stimuli in high-level semantic features. This also overcomes the limitation of existing cross-subject methods that optimize only between pairs of subjects. BrainCanvas also maintains strong performance in other metrics, showing high similarity to visual stimuli in object contours and image structure. Visualization results are presented in Figure 5. By combining ImageSet Diffusion's ability to mitigate data scarcity with the Domain-Invariant Encoder's effectiveness in cross-subject feature learning, BrainCanvas effectively restores accurate semantics. These results demonstrate that BrainCanvas not only matches or exceeds single-subject methods but also improves cross-subject brain decoding through domain-invariant representation learning, significantly reducing the dependence on extensive subject-specific training data.

4.4 Data Scarcity and New Subject Adaptation

Zero-Shot Cross-Subject Transfer. To validate the cross-subject generalization capability of our Domain-Invariant Encoder, we first conduct a Zero-Shot experiment. Specifically, we use the model pre-trained on Subjects 1, 2, and 5 to perform inference directly on Subject 7 without any gradient updates or fine-tuning. As shown in Table 2, even under this challenging Zero-Shot setting where the model has never encountered any data from Subject 7, BrainCanvas achieves a CLIP Score of 80.1% and an Inception Score of 77.3%. This result strongly validates that our Domain-Invariant Encoder effectively learns subject-invariant feature representations and aligns the feature spaces across different subjects. Furthermore, with minimal fine-tuning using only 500 samples, performance improves substantially to 89.2% CLIP Score, demonstrating the model's efficient adaptation capability. Fine-tuning with full data achieves state-of-the-art performance, with a CLIP Score of 96.5%.

Few-Shot Adaptation Analysis. To evaluate BrainCanvas's cross-subject transfer performance, we conducted quantitative and qualitative analyses under data-scarce scenarios, evaluating reconstruction quality and generalization potential. We employed a stratified sampling strategy, utilizing 500, 1500, and 4000 samples from the dataset for adaptation tests on new individuals. Pre-training utilized three subjects' data, while an independent subject's data validated adaptation validation, as detailed in Table 3. BrainCanvas enhanced semantic diversity through Image Diffusion augmentation, generating supplemental 512×512 imagesets (marked with * in the table). The results clearly show that as the number of training samples increases, the model's performance on new subject gradually approaches the fully pre-trained benchmark level. Moreover, our method significantly outperforms existing approaches across all evaluation metrics and data scales. Notably, BrainCanvas maintains high reconstruction quality even with a minimal dataset of 500 samples. The transfer results are illustrated in Figure 6, validating BrainCanvas's effectiveness in cross-subject transfer. The model successfully maps fMRI signal features from source subjects to the target subject's feature space, enabling accurate reconstruction of the target subject's brain activity images.

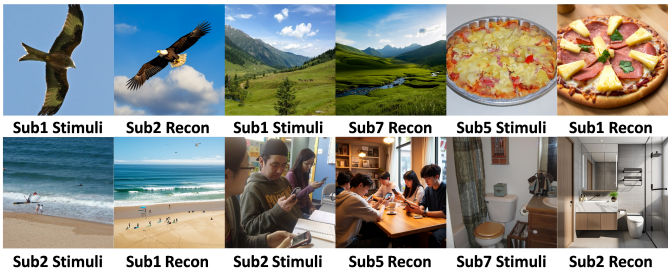


Fig. 6 Cross-subject brain signals reconstruction results.

4.5 Source Domain Diversity Analysis.

To more rigorously validate that the Domain-Invariant Encoder genuinely learns universal features of neural activity rather than merely memorizing data distributions, we conducted supplementary experiments examining the effect of source domain diversity on generalizability. Specifically, we progressively increased the number of source subjects used for pre-training while evaluating performance on a fixed target subject (Subject 7) under a 500-shot fine-tuning setting. The results are presented in Table 4.

As the number of source domain subjects increases, model performance on the target subject demonstrates consistent and steady improvement across all evaluation metrics. The transition from training from scratch to leveraging a single source subject (S1) yields substantial gains, with Inception scores improving from 63.2% to 78.5% and CLIP scores from 66.5% to 81.2%. Further incorporating additional source subjects (S1, S2 and subsequently S1, S2, S5) continues to enhance performance, ultimately achieving 87.1% Inception and 89.2% CLIP scores with three source subjects. These findings provide strong evidence that BrainCanvas can effectively extract and leverage shared features across different individuals. Moreover, as the source domain becomes increasingly diverse, the learned feature space more closely approximates the ideal subject-invariant state, thereby conferring enhanced generalization potential. This confirms that the Domain-Invariant Encoder learns genuinely universal neural representations rather than overfitting to subject-specific patterns.

Table 1 Quantitative comparison of fMRI-to-image reconstruction models. Performance is evaluated for both single-subject and cross-subject methods, with BrainCanvas and MindBridge representing the latter. In the results, boldface indicates the best performance and underlining the second.

Type	Method	High-Level				Low-Level			
		Incep ↑	CLIP ↑	EffNet-B ↓	SwAV ↓	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑
Single Subject	Takagi. [24] *[CVPR'23]*	76.0%	77.0%	–	–	–	–	83.0%	83.0%
	Mind-Reader [50]*[NeurIPS'22]*	78.2%	–	–	–	–	–	–	–
	Mind-Vis [51] *[CVPR'23]*	78.8%	76.2%	0.854	0.491	0.080	0.220	72.1%	83.2%
	Gu. [42] *[MIDL'24]*	–	–	0.862	0.465	0.150	0.325	–	–
	Brain-Diffuser [6]*[SciRep'23]*	87.2%	91.5%	0.775	0.423	0.254	0.356	94.2%	96.2%
	MindEye [25]*[NeurIPS'24]*	93.8%	93.6%	**0.645**	0.377	**0.309**	0.323	92.8%	**96.9%**
	MindBridge [26]*[CVPR'24]*	92.2%	94.3%	0.713	0.413	0.148	0.259	86.9%	95.3%
	BrainCanvas (Ours)	**94.8%**	**95.6%**	0.647	**0.352**	0.291	**0.386**	**95.2%**	96.8%
Cross Subject	MindBridge [26]*[CVPR'24]*	92.4%	94.7%	0.712	0.418	0.151	0.263	87.7%	95.5%
	Wills Aligner [36]*[AAAI'25]*	94.3%	94.8%	0.649	0.373	0.271	0.328	95.8%	98.0%
	BrainGuard [37]*[AAAI'25]*	94.9%	96.4%	0.624	0.353	0.313	0.330	94.7%	97.8%
	MindEye2 [34]*[ICML'24]*	95.4%	93.0%	0.619	0.344	0.322	**0.427**	95.8%	98.1%
	UMBRAE [35]*[ECCV'24]*	91.7%	93.5%	0.700	0.393	0.283	0.341	95.5%	97.0%
	BrainCanvas (Ours)	**96.0%**	**96.5%**	**0.615**	**0.337**	**0.324**	0.401	**96.3%**	**98.4%**

Table 2 Zero-Shot and Fine-Tuned Performance on New Subject (Subject 7). The model is pre-trained on Subjects 1, 2, and 5. Zero-Shot: direct inference without any gradient updates; Fine-Tuned (500): fine-tuning with 500 samples; Fine-Tuned (Full): fine-tuning with full training data.

Setting	Epochs	High-Level				Low-Level			
		Incep ↑	CLIP ↑	EffNet-B ↓	SwAV ↓	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑
Zero-Shot	0	77.3%	80.1%	0.865	0.552	0.090	0.182	70.5%	81.8%
Fine-Tuned (500)	200	87.1%	89.2%	0.774	0.461	0.138	0.264	84.4%	92.6%
Fine-Tuned (Full)	200	**96.0%**	**96.5%**	**0.615**	**0.337**	**0.324**	**0.401**	**96.3%**	**98.4%**

5 Ablation Study

5.1 Ablation Study on Generated Imageset Scale

We systematically evaluated how the scale of the generated imageset impacts model performance. As shown in Table 5, expanding the imageset significantly enhances reconstruction quality. This improvement stems from the augmented data providing richer and more diverse features, which enables the model to capture more comprehensive neural patterns and thus boosts generalization. However, further expanding the generated imageset beyond three times the original size does not yield additional performance gains. Although the increased data partially alleviates scarcity by enlarging the training set, an excessive amount of generated imageset may introduce redundant or noisy information, failing to contribute meaningful new features and potentially causing unnecessary interference.

5.2 Ablation Study of Pretrain Loss

Within the BrainCanvas framework, fMRI signals are mapped to the CLIP embedding space to guide the diffusion model in generating reconstructed images. We optimize this process using multiple loss functions, including SoftCLIP loss, MSE loss, reconstruction loss, and domain adversarial loss to ensure semantic consistency across subjects. To monitor the degree of subject-invariance in the learned features, we introduce Domain Classification Accuracy as a diagnostic metric. A 4-class classifier is trained to predict subject identity from the features, where the ideal value of 25% corresponds to random chance, indicating complete removal of subject-specific information.

Table 6 demonstrates that different loss combinations significantly impact both reconstruction quality and domain invariance. Using only SoftCLIP loss achieves high SSIM scores but overlooks high-level semantic information, with Domain Accuracy at 87.5%. Combining SoftCLIP and MSE losses reduces Domain Accuracy to 78.5% while substantially improving high-level metrics. A critical observation emerges when adding the self-supervised reconstruction loss $\mathcal{L}_{\text{rec}}$. Although it improves reconstruction fidelity such as SSIM from 0.395 to 0.402, it simultaneously increases Domain Accuracy from 78.5% to 89.4%, as the reconstruction task forces the encoder to retain subject-specific neural patterns. This provides a clear justification for adversarial training. Upon introducing $\mathcal{L}_{\text{adv}}$, Domain Accuracy drops dramatically to 33.8%, approaching the 25% chance level, demonstrating that adversarial training effectively "unlearns" subject identity information from the feature representations. This removal of subject-specific bias is accompanied by further improvements across all high-level metrics, confirming that the cross-subject performance gains are fundamentally attributable to the domain alignment mechanism rather than merely the reconstruction constraint.

Table 3 Quantitative evaluation of cross-subject adaptation capabilities in data-constrained settings. BrainCanvas(Ours) is finetuned on subject 7 from model pretrained on subject 1, 2 and 5. Results demonstrate the model's transfer learning efficiency and generalization capacity under varying data availability conditions.

Method	Data	High-Level				Low-Level			
		Incep ↑	CLIP ↑	EffNet-B ↓	SwAV ↓	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑
MindBridge	500	82.3%	86.7%	0.840	0.521	0.112	0.229	79.6%	89.0%
BrainCanvas	**500***	**87.1%**	**89.2%**	**0.774**	**0.461**	**0.138**	**0.264**	**84.4%**	**92.6%**
MindBridge	1500	85.8%	91.0%	0.796	0.485	0.140	0.250	84.6%	92.6%
BrainCanvas	**1500***	**91.3%**	**92.7%**	**0.705**	**0.395**	**0.192**	**0.319**	**89.8%**	**95.0%**
MindBridge	4000	88.9%	92.5%	0.765	0.458	0.156	0.258	85.7%	94.1%
BrainCanvas	**4000***	**95.2%**	**95.1%**	**0.638**	**0.349**	**0.301**	**0.388**	**95.4%**	**98.0%**

Table 4 Impact of Source Subject Diversity on Target Subject 7 (500-shot fine-tuning).

Subjects	High-Level				Low-Level			
	Incep ↑	CLIP ↑	EffNet-B ↓	SwAV ↓	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑
Scratch	63.2%	66.5%	0.962	0.625	0.045	0.132	64.5%	75.3%
S1	78.5%	81.2%	0.865	0.540	0.090	0.215	77.2%	87.5%
S1, S2	84.8%	85.9%	0.792	0.485	0.126	0.248	81.6%	90.8%
S1, S2, S5	**87.1%**	**89.2%**	**0.774**	**0.461**	**0.138**	**0.264**	**84.4%**	**92.6%**

5.3 Ablation Study of Different Generating Imageset Resolutions

We conducted an ablation study to assess the impact of image resolution on model performance across various metrics. As depicted in Table 7, increasing the resolution from 128 to 512 consistently enhances performance in most high-level and low-level metrics. These results demonstrate that higher generating imageset resolutions contribute to more accurate and detailed visual resolutions, confirming the positive effect of increased resolution on overall model performance.

5.4 Ablation Study of Components in the Imageset Diffusion Model

We conduct an ablation study to meticulously evaluate the contribution of each component within our proposed Imageset Diffusion framework. The results are presented in Table 8. The 'Captions only' utilizes solely the descriptive image captions as textual prompts to guide the model in generating imagesets. Under the 'Class labels only' setting, only the object class labels present in the images serve as textual prompts. This configuration, as evidenced by the results, leads to a significant degradation in performance metrics. The 'Caption + class labels' approach investigates the utility of augmenting captions by appending the respective class labels, thereby forming a richer textual prompt. Finally, 'Total' denotes the incorporation of Perceptual Position in addition to the combined caption and class labels. Our ablation experiments systematically demonstrate that all aforementioned operations contribute positively to the overall performance gains achieved by our proposed method.

5.5 Fair Comparison Under Data Augmentation

To ensure a fair comparison and demonstrate that performance improvements stem from our proposed architecture rather than solely from increased data volume, we conducted an additional experiment where MindBridge was retrained using the identical augmented imageset generated through our Imageset Diffusion strategy.

As shown in Table 9, utilizing augmented data improves the performance of MindBridge across all metrics, which validates the effectiveness of our proposed Imageset Diffusion strategy as a general-purpose data augmentation technique applicable beyond our own framework. However, even under this data-fair condition, BrainCanvas still significantly outperforms the augmented MindBridge.

5.6 Hyperparameter Sensitivity Analysis

In our BrainCanvas framework, three core hyperparameters govern the optimization process as defined in Equation 11: the adversarial loss weight λ, the image alignment weight α, and the text alignment weight β. To validate the rationale behind our selected values ($\lambda = 0.1$, $\alpha = 1$, $\beta = 10^4$), we conducted a univariate sensitivity analysis on the full cross-subject dataset, varying one parameter at a time while holding the remaining parameters fixed at their default values. The results are presented in Table 10.

The sensitivity analysis reveals several important insights. First, the adversarial loss weight λ plays a critical role in cross-subject generalization. Removing the adversarial component entirely ($\lambda = 0$) leads to a notable performance drop, confirming the necessity of the Domain-Invariant Encoder for learning subject-invariant representations. However, setting λ too high ($\lambda = 1.0$) causes the adversarial training to dominate and destabilize the learning process, resulting in degraded performance across all metrics.

Second, the image alignment weight α controls the trade-off between low-level visual fidelity and high-level semantic flexibility. Increasing α to 10 improves low-level metrics such as SSIM and PixCorr , but slightly compromises high-level semantic metrics. Conversely, reducing α to 0.1 weakens low-level reconstruction quality. Moreover,

Table 5 Impact of generated imageset scale on model performance. The training dataset was augmented by factors of double, triple, and quadruple using our semantic diversification method. All generated images were 512×512 pixels.

Data Scale	High-Level				Low-Level			
	Incep ↑	CLIP ↑	EffNet-B ↓	SwAV ↓	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑
Original	93.9%	95.9%	0.650	0.374	0.288	0.354	94.6%	97.6%
Double	95.3%	96.2%	0.637	0.349	0.301	0.387	96.1%	98.0%
Triple	**96.0%**	**96.5%**	0.615	**0.337**	0.324	**0.401**	**96.3%**	**98.4%**
Quadruple	95.7%	96.3%	**0.613**	0.344	**0.328**	0.398	96.1%	98.3%

Table 6 Comprehensive Ablation Study on Loss Functions. We merge data to analyze the impact of individual components. Note that the adversarial loss ($\mathcal{L}_{\text{adv}}$) is critical for reducing Domain Accuracy (improving anonymization) while maintaining high generation quality.

Loss Components				Domain	High-Level				Low-Level			
$\mathcal{L}_{\text{SoftCLIP}}$	$\mathcal{L}_{\text{MSE}}$	$\mathcal{L}_{\text{rec}}$	$\mathcal{L}_{\text{adv}}$	Acc ↓	Incep ↑	CLIP ↑	EffNet-B ↓	SwAV ↓	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑
✓				87.5%	85.3%	84.9%	0.765	0.450	0.207	**0.453**	86.6%	90.6%
	✓			86.2%	85.5%	86.4%	0.742	0.465	0.184	0.311	88.1%	90.8%
✓	✓			78.5%	95.5%	95.8%	0.637	0.346	0.286	0.395	96.4%	98.2%
✓	✓	✓		89.4%	95.8%	96.1%	0.628	0.340	0.308	0.402	96.9%	98.4%
✓	✓	✓	✓	**33.8%**	**96.0%**	**96.7%**	**0.616**	**0.334**	**0.328**	0.410	**97.1%**	**98.7%**



Fig. 7 More cross-subject reconstructions of BrainCanvas on subject 1, 2, 5 and 7.



Fig. 8 Comparison of Similar Image Reconstructions Across Datasets: Left: NSD Dataset, Right: GOD Dataset.

the text alignment weight β exhibits a similar trade-off pattern. A high semantic weight ($\beta = 10^5$) yields the best CLIP score but degrades structural fidelity metrics. A low weight ($\beta = 10^2$) improves pixel-level metrics but sacrifices semantic consistency. These results demonstrate that our default configuration ($\lambda = 0.1$, $\alpha = 1$, $\beta = 10^4$) achieves an optimal balance between semantic consistency and visual accuracy across all evaluation metrics.

■ 6 More Reconstruction Results

To further demonstrate the cross-subject generalization capabilities of BrainCanvas, Figure 7 presents additional visual results of image reconstruction from brain signals. The consistency and remarkable fidelity of these outcomes when applied to various individuals effectively showcase our model's robust performance. These results underscore BrainCanvas's ability to maintain excellent visual representation quality across multiple subjects, affirming its strong generalization potential for diverse neural patterns.

■ 7 Comparative Analysis of Visual Reconstruction Across Datasets

To comprehensively demonstrate the visual reconstruction capabilities of our BrainCanvas, we compare it with two models based on the Generic Object Decoding Dataset (GOD) [20]: MinD-Vis [51] and IC-GAN [22]. As shown in Figure 8, BrainCanvas significantly outperforms these methods in reconstructing the original visual stimuli. One possible reason for this substantial difference is that the NSD dataset we used offers richer content diversity and more complex scene structures compared to the GOD dataset. Moreover, our proposed Semantic Diversified Imageset Diffusion method introduces more abundant semantic information and image diversity during model training. This indirectly demonstrates the critical role of incorporating diversified semantic information and enhancing dataset quality in visual reconstruction tasks.

Table 7 Comparison of Performance Metrics Across Image Resolutions: 128, 256, and 512

Resolution	High-Level				Low-Level			
	Incep ↑	CLIP ↑	EffNet-B ↓	SwAV ↓	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑
128	93.9%	95.6%	0.624	0.351	0.285	0.371	94.1%	97.2%
256	95.3%	96.2%	**0.614**	0.340	0.306	0.397	95.4%	97.8%
512	**96.0%**	**96.5%**	0.615	**0.337**	**0.324**	**0.401**	**96.3%**	**98.4%**

Table 8 Ablation Study of Components in Imageset Diffusion. C_{Cap}: caption loss, C_{Cls}: class loss, $C_{\text{Cap+Cls}}$: joint caption–class loss, $C_{\text{Cap+Cls+PP}}$: total loss with perceptual position.

Loss			High-Level				Low-Level			
C_{Cap}	C_{Cls}	C_{PP}	Incep ↑	CLIP ↑	EffNet-B ↓	SwAV ↓	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑
✓			93.7%	95.2%	0.637	0.373	0.297	0.383	95.8%	98.0%
	✓		79.4%	81.1%	0.718	0.386	0.202	0.310	86.8%	86.0%
✓	✓		95.1%	95.9%	0.624	0.342	0.321	0.393	**96.3%**	98.2%
✓	✓	✓	**96.0%**	**96.5%**	**0.615**	**0.337**	**0.324**	**0.401**	**96.3%**	**98.4%**

8 Text Prompt Selection Details

In this section, we investigate methods for selecting and processing text prompts under different contexts. For simple class prompts, we construct the text prompt as "Original Caption; [Class Name]," where "[Class Name]" denotes the target category. In this setup, the cross-attention map primarily focuses on "[Class Name]," thereby reinforcing the model's attention to features specific to that category. For multiple class prompts, we use the format "Original Caption; [Class Name 1] [Class Name 2] ...," enabling the cross-attention mechanism to aggregate information from multiple categories. In cases of compound category names like "coffee table," we need to synthesize the attention information of each word across all positions, effectively integrating the semantic information of the compound category and ensuring the model's accurate understanding of complex categories.

Through analyzing the self-attention heatmaps in Figure 9, we observe that the heatmaps display attention scores across different spatial locations. Each heatmap highlights regions highly relevant to specific categories in the input, validating the self-attention mechanism's ability to focus on semantically important regions. This proficiency further facilitates the generation of refined cross-attention maps. The proposed flexible text prompt processing strategy ensures that the generated imageset are more faithful to the provided text prompts.

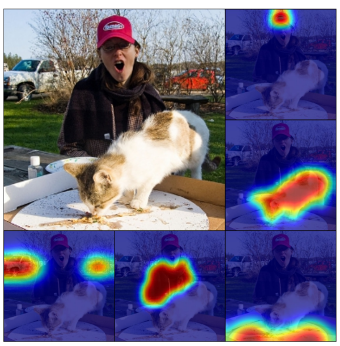


Fig. 9 Heatmap visualizations of self-attention maps used to refine cross-attention maps, generating higher-quality and more accurate imageset.

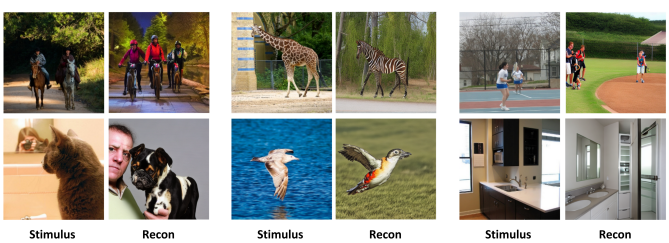


Fig. 10 Representative failure cases in cross-subject reconstruction. The model occasionally produces semantically related but visually distinct objects (e.g., giraffe→zebra, tennis ball→baseball) and may fail to preserve spatial arrangements in complex scenes.

9 Limitations and Failure Case Analysis

Although BrainCanvas achieves state-of-the-art performance in cross-subject reconstruction tasks, we conducted a systematic analysis to identify challenging scenarios and model limitations. As shown in Figure 10, by comparing visual stimuli with reconstructed images, we observe that the model occasionally produces objects that are semantically related but visually distinct from the ground truth. For instance, a giraffe may be reconstructed as a zebra, horseback riding as cycling, and a tennis ball as a baseball. This phenomenon likely arises because the Domain-Invariant Encoder maps fMRI signals into a semantic space where conceptually similar categories are positioned in close proximity. Under low signal-to-noise ratio conditions, feature embeddings may drift into the vector domain of neighboring categories, thereby guiding the diffusion model to generate semantically synonymous but visually different objects. Additionally, in scenes with complex layouts or multiple small objects, reconstruction results sometimes fail to preserve the relative spatial positions of objects, such as the arrangement of fixtures in a bathroom.

Although we introduce the Category-aware Localization Module

Table 9 Performance Comparison: Effect of Data Augmentation and Model Architecture. Results demonstrate that while Imageset Diffusion improves MindBridge, BrainCanvas's architectural design provides additional significant gains.

Method	Data	High-Level				Low-Level			
		Incep↑	CLIP↑	EffNet-B↓	SwAV↓	PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑
MindBridge	Original	92.4%	94.7%	0.712	0.418	0.151	0.263	87.7%	95.5%
	Augmented	94.8%	96.2%	0.675	0.380	0.220	0.325	92.2%	96.4%
BrainCanvas	Augmented	**96.0%**	**96.5%**	**0.615**	**0.337**	**0.324**	**0.401**	**96.3%**	**98.4%**

Table 10 Hyperparameter Sensitivity Analysis. We analyze the impact of the adversarial weight (λ), visual weight (α), and semantic weight (β) relative to the default configuration.

Parameter	Value	High-Level				Low-Level			
		Incep ↑	CLIP ↑	EffNet-B ↓	SwAV ↓	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑
Default	-	96.0%	96.5%	0.615	0.337	0.324	0.401	96.3%	98.4%
λ (Adv.)	0.0	94.1%	94.6%	0.662	0.372	0.306	0.382	94.8%	97.5%
	0.5	94.8%	95.2%	0.641	0.355	0.312	0.389	95.4%	97.8%
	1.0	92.3%	92.9%	0.704	0.404	0.293	0.357	93.3%	96.5%
α (Image)	0.1	95.2%	95.8%	0.631	0.348	0.295	0.368	94.6%	97.6%
	10	95.0%	95.4%	0.645	0.361	0.332	0.409	96.0%	98.1%
β (Text)	10^2	94.6%	94.0%	0.654	0.372	0.331	0.407	95.8%	98.1%
	10^5	95.4%	96.9%	0.635	0.348	0.306	0.382	95.1%	97.6%

and Perceptual Position Map to enhance spatial guidance, structural deviations in complex scenes may still occur due to the inherent spatial resolution limitations of fMRI signals and the stochastic nature of the diffusion generation process. It is worth noting that these failure cases are not unique to BrainCanvas, but rather reflect common challenges in the fMRI visual decoding field. Non-invasive fMRI technology inherently suffers from low signal-to-noise ratio and limited spatiotemporal resolution, meaning that high-frequency pixel-level details are often lost in the neural signals. Therefore, certain semantic ambiguities or detail losses reflect the inherent limitations of the fMRI data modality rather than architectural deficiencies. Our model has demonstrated superior robustness compared to existing methods in extracting consistent semantics from such noisy data.

■ 10 Discussion and Conclusion

In this study, we present BrainCanvas, a novel framework addressing data scarcity and cross-subject generalization in brain signal decoding. BrainCanvas employs semantic diversification imageSet diffusion to generate multiple semantically consistent images from identical neural inputs, mitigating data limitations. An adversarial mechanism ensures consistent feature distributions across subjects, enabling the model to learn universal, subject-invariant representations and enhancing generalization. NSD dataset evaluations demonstrate superior performance, with ablation studies validating component efficacy. However, Imageset Diffusion may exhibit limitations in generating intricate scenes; excessive class specifications can yield chaotic or incoherent outputs, necessitating subsequent manual screening. Future directions include extending to EEG modalities, neural video reconstruction, and 3D neural rendering. Additionally, as neural technologies advance, developing robust data privacy and ethical guidelines is essential, requiring transparency from data collectors and collaboration among researchers to establish standard protocols.

■ Appendixes

Supplementary results for new subject adaptation under limited data conditions, are detailed in Table 11, Table 12, and Table 13, respectively. The results show that our method consistently yields superior performance compared to the other method.

■ Acknowledgement

This work was supported by the STI 2030 Major Projects of China under Grant No. 2021ZD0200400 and the National Natural Science Foundation of China under Grant No. 62506336.

■ Competing Interest

The authors declare that they have no competing interests or financial conflicts to disclose.

■ Data and Code Availability

The source code for BrainCanvas are publicly available at `https://github.com/ori0n112/BrainCanvas`. The NSD dataset used in this work is accessible at `https://naturalscenesdataset.org`.

Table 11 Quantitative evaluation of cross-subject adaptation capabilities in data-constrained settings. BrainCanvas is finetuned on subject 1 from model pretrained on subject 2, 5, 7.

Method	Data	High-Level				Low-Level			
		Incep ↑	CLIP ↑	EffNet-B ↓	SwAV ↓	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑
MindBridge	500	82.8%	89.4%	0.831	0.500	0.142	0.263	84.4%	91.4%
BrainCanvas	**500***	**85.2%**	**91.7%**	**0.755**	**0.445**	**0.230**	**0.365**	**89.8%**	**94.5%**
MindBridge	1500	87.7%	92.5%	0.784	0.460	0.161	0.275	87.2%	94.2%
BrainCanvas	**1500***	**90.0%**	**94.8%**	**0.705**	**0.405**	**0.252**	**0.378**	**92.5%**	**97.1%**
MindBridge	4000	91.5%	93.9%	0.747	0.436	0.157	0.275	88.1%	95.5%
BrainCanvas	**4000***	**95.6%**	**96.1%**	**0.622**	**0.345**	**0.325**	**0.403**	**97.0%**	**98.8%**

Table 12 Quantitative evaluation of cross-subject adaptation capabilities in data-constrained settings. BrainCanvas is finetuned on subject 2 from model pretrained on subject 1, 5, 7.

Method	Data	High-Level				Low-Level			
		Incep ↑	CLIP ↑	EffNet-B ↓	SwAV ↓	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑
MindBridge	500	83.0%	87.2%	0.833	0.501	0.122	0.265	83.2%	91.0%
BrainCanvas	**500***	**86.6%**	**92.4%**	**0.748**	**0.440**	**0.234**	**0.378**	**90.7%**	**95.4%**
MindBridge	1500	88.2%	91.7%	0.780	0.459	0.143	0.273	86.6%	94.2%
BrainCanvas	**1500***	**94.3%**	**94.2%**	**0.640**	**0.373**	**0.253**	**0.382**	**92.8%**	**96.7%**
MindBridge	4000	90.1%	93.0%	0.745	0.436	0.150	0.272	88.4%	95.5%
BrainCanvas	**4000***	**95.9%**	**96.4%**	**0.617**	**0.330**	**0.320**	**0.405**	**95.8%**	**98.3%**

Table 13 Quantitative evaluation of cross-subject adaptation capabilities in data-constrained settings. BrainCanvas is finetuned on subject 5 from model pretrained on subject 1, 2, 7.

Method	Data	High-Level				Low-Level			
		Incep ↑	CLIP ↑	EffNet-B ↓	SwAV ↓	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑
MindBridge	500	86.2%	91.2%	0.794	0.477	0.124	0.255	83.1%	91.9%
BrainCanvas	**500***	**89.0%**	**93.4%**	**0.715**	**0.420**	**0.235**	**0.363**	**90.5%**	**95.7%**
MindBridge	1500	89.8%	93.5%	0.746	0.440	0.135	0.262	86.1%	93.8%
BrainCanvas	**1500***	**94.0%**	**95.7%**	**0.635**	**0.365**	**0.260**	**0.377**	**91.5%**	**96.8%**
MindBridge	4000	91.9%	95.0%	0.712	0.418	0.150	0.267	87.9%	95.6%
BrainCanvas	**4000***	**96.2%**	**97.2%**	**0.615**	**0.333**	**0.325**	**0.400**	**95.2%**	**98.4%**

References

[1] Du B, Cheng X, Duan Y, Ning H. fmri brain decoding and its applications in brain–computer interface: A survey. Brain Sciences, 2022, 12(2): 228

[2] Teng C, Kravitz D J. Visual working memory directly alters perception. Nature human behaviour, 2019, 3(8): 827–836

[3] Heekeren H R, Marrett S, Bandettini P A, Ungerleider L G. A general mechanism for perceptual decision-making in the human brain. Nature, 2004, 431(7010): 859–862

[4] Yoo P E, John S E, Farquharson S, Cleary J O, Wong Y T, Ng A, Mulcahy C B, Grayden D B, Ordidge R J, Opie N L, others . 7t-fmri: Faster temporal resolution yields optimal bold sensitivity for functional network imaging specifically at high spatial resolution. Neuroimage, 2018, 164: 214–229

[5] Lu Y, Du C, Zhou Q, Wang D, He H. Minddiffuser: Controlled image reconstruction from human brain activity with semantic and structural diffusion. In: Proceedings of the 31st ACM International Conference on Multimedia. 2023, 5899–5908

[6] Ozcelik F, VanRullen R. Natural scene reconstruction from fmri signals using generative latent diffusion. Scientific Reports, 2023, 13(1): 15666

[7] 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. 2022, 10684–10695

[8] Wang T, Zhang T, Zhang B, Ouyang H, Chen D, Chen Q, Wen F. Pretraining is all you need for image-to-image translation. arXiv preprint arXiv:2205.12952, 2022

[9] 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, 2020, 63(11): 139–144

[10] Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, others . Learning transferable visual models from natural language supervision. In: International conference on machine learning. 2021, 8748–8763

[11] Yin W, Li L, Wu F X. Deep learning for brain disorder diagnosis based on fmri images. Neurocomputing, 2022, 469: 332–345

[12] Yan W, Qu G, Hu W, Abrol A, Cai B, Qiao C, Plis S M, Wang Y P, Sui J, Calhoun V D. Deep learning in neuroimaging: Promises and challenges. IEEE Signal Processing Magazine, 2022, 39(2): 87–98

[13] Wei C S, Keller C J, Li J, Lin Y P, Nakanishi M, Wagner J, Wu W, Zhang Y, Jung T P. Inter-and intra-subject variability in brain imaging

and decoding, 2021

[14] Edelman B J, Zhang S, Schalk G, Brunner P, Müller-Putz G, Guan C, He B. Non-invasive brain-computer interfaces: State of the art and trends. IEEE Reviews in Biomedical Engineering, 2024

[15] Stansbury D E, Naselaris T, Gallant J L. Natural scene statistics account for the representation of scene categories in human visual cortex. Neuron, 2013, 79(5): 1025–1034

[16] Wardle S G, Kriegeskorte N, Grootswagers T, Khaligh-Razavi S M, Carlson T A. Perceptual similarity of visual patterns predicts dynamic neural activation patterns measured with meg. Neuroimage, 2016, 132: 59–70

[17] Stojanov P, Li Z, Gong M, Cai R, Carbonell J, Zhang K. Domain adaptation with invariant representation learning: What transformations to learn? Advances in Neural Information Processing Systems, 2021, 34: 24791–24803

[18] Han K, Wen H, Shi J, Lu K H, Zhang Y, Fu D, Liu Z. Variational autoencoder: An unsupervised model for encoding and decoding fmri activity in visual cortex. NeuroImage, 2019, 198: 125–136

[19] Livezey J A, Glaser J I. Deep learning approaches for neural decoding across architectures and recording modalities. Briefings in bioinformatics, 2021, 22(2): 1577–1591

[20] Horikawa T, Kamitani Y. Generic decoding of seen and imagined objects using hierarchical visual features. Nature communications, 2017, 8(1): 15037

[21] Donahue J, Simonyan K. Large scale adversarial representation learning. Advances in neural information processing systems, 2019, 32

[22] Casanova A, Careil M, Verbeek J, Drozdzal M, Romero Soriano A. Instance-conditioned gan. Advances in Neural Information Processing Systems, 2021, 34: 27517–27529

[23] Ma X, Fang G, Wang X. Deepcache: Accelerating diffusion models for free. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024, 15762–15772

[24] Takagi Y, Nishimoto S. 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. 2023, 14453–14463

[25] Scotti P, Banerjee A, Goode J, Shabalin S, Nguyen A, Dempster A, Verlinde N, Yundler E, Weisberg D, Norman K, others . Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. Advances in Neural Information Processing Systems, 2024, 36

[26] Wang S, Liu S, Tan Z, Wang X. Mindbridge: A cross-subject brain decoding framework. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024, 11333–11342

[27] Wolleb J, Sandkühler R, Bieder F, Valmaggia P, Cattin P C. Diffusion models for implicit image segmentation ensembles. In: International Conference on Medical Imaging with Deep Learning. 2022, 1336–1348

[28] Gu Z, Chen H, Xu Z. Diffusioninst: Diffusion model for instance segmentation. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2024, 2730–2734

[29] Sarıyıldız M B, Alahari K, Larlus D, Kalantidis Y. Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023, 8011–8021

[30] He R, Sun S, Yu X, Xue C, Zhang W, Torr P, Bai S, Qi X. Is synthetic data from generative models ready for image recognition? arXiv preprint arXiv:2210.07574, 2022

[31] Azizi S, Kornblith S, Saharia C, Norouzi M, Fleet D J. Synthetic data from diffusion models improves imagenet classification. arXiv preprint arXiv:2304.08466, 2023

[32] Deng J, Dong W, Socher R, Li L J, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. 2009, 248–255

[33] Nguyen Q, Vu T, Tran A, Nguyen K. Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. Advances in Neural Information Processing Systems, 2024, 36

[34] Scotti P S, Tripathy M, Villanueva C K T, Kneeland R, Chen T, Narang A, Santhirasegaran C, Xu J, Naselaris T, Norman K A, others . Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. arXiv preprint arXiv:2403.11207, 2024

[35] Xia W, Charette d R, Oztireli C, Xue J H. Umbrae: Unified multimodal brain decoding. In: European Conference on Computer Vision. 2024, 242–259

[36] Bao G, Zhang Q, Gong Z, Zhou J, Fan W, Yi K, Naseem U, Hu L, Miao D. Wills aligner: Multi-subject collaborative brain visual decoding. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2025, 14194–14202

[37] Tian Z, Quan R, Ma F, Zhan K, Yang Y. Brainguard: Privacy-preserving multisubject image reconstructions from brain activities. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2025, 14414–14422

[38] Ganin Y, Lempitsky V. Unsupervised domain adaptation by backpropagation. In: International conference on machine learning. 2015, 1180–1189

[39] Xu X, Wang Z, Zhang G, Wang K, Shi H. Versatile diffusion: Text, images and variations all in one diffusion model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023, 7754–7765

[40] Allen E J, St-Yves G, Wu Y, Breedlove J L, Prince J S, Dowdle L T, Nau M, Caron B, Pestilli F, Charest I, others . A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. Nature neuroscience, 2022, 25(1): 116–126

[41] Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick C L. Microsoft coco: Common objects in context. in david fleet tomas pajdla bernt schiele and tinne tuytelaars editors. In: Computer Vision–ECCV 2014. 2014, 740–755

[42] Gu Z, Jamison K, Kuceyeski A, Sabuncu M R. Decoding natural image stimuli from fmri data with a surface-based convolutional network. In: Medical Imaging with Deep Learning. 2024, 107–118

[43] Mai W, Zhang Z. Unibrain: Unify image reconstruction and

captioning all in one diffusion model from human brain activity. arXiv preprint arXiv:2308.07428, 2023

[44] Xia W, Charette d R, Oztireli C, Xue J H. Dream: Visual decoding from reversing human visual system. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2024, 8226–8235

[45] Wang Z, Bovik A C, Sheikh H R, Simoncelli E P. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 2004, 13(4): 600–612

[46] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 2012, 25

[47] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, 2818–2826

[48] Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. 2019, 6105–6114

[49] Caron M, Misra I, Mairal J, Goyal P, Bojanowski P, Joulin A. Unsupervised learning of visual features by contrasting cluster assignments. Advances in neural information processing systems, 2020, 33: 9912–9924

[50] Lin S, Sprague T, Singh A K. Mind reader: Reconstructing complex images from brain activities. Advances in Neural Information Processing Systems, 2022, 35: 29624–29636

[51] 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. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023, 22710–22720