MindPainter: Efficient Brain-Conditioned Painting of Natural Images via Cross-Modal Self-Supervised Learning

Muzhou Yu[1*], **Shuyun Lin**[2*], **Hongwei Yan**[2], **Kaisheng Ma**[2†]

[1]Xi'an Jiaotong University
[2]Tsinghua University
muzhou9999@stu.xjtu.edu.cn, {linshuyu21, yanhw22}@mails.tsinghua.edu.cn, kaisheng@mail.tsinghua.edu.cn

Abstract

Despite significant advancements in image and text conditional image editing, the exploration of using brain signals, which are more direct and personalized to reflect user intentions, remains limited. An intuitive method is to convert implicit brain signals into explicit representations such as images, which can then serve as prompts for editing. However, such two-stage method suffers from low inference efficiency, inaccurate brain interpretation, and unnatural editing results. In this paper, we apply brain signals of visual perception as prompts and propose a cross-modal self-supervised learning for natural image painting (*MindPainter*). This method achieves efficient and natural brain-conditioned image editing in a straightforward manner. MindPainter is trained for reconstruction from masked images directly with pseudo-brain signals, which is simulated by the proposed Pseudo Brain Generator. It facilitates efficient cross-modal integration. The proposed Brain Adapter further eliminates the gap in implicit space between modalities, ensuring accurate semantic interpretation of brain signals and coherent consolidation. Besides, the designed Multi-Mask Generation Policy enhances the generalization, realizing high-quality editing in various painting scenarios, including inpainting and outpainting. To the best of our knowledge, MindPainter is the first work to achieve efficient brain-conditioned image painting, providing potential for direct brain control in creative AI. The code and the link to the extended version will be available on GitHub.

Introduction

Image editing (Oh et al. 2001; Kawar et al. 2023; Wang et al. 2024; Suvorov et al. 2022a) is a powerful tool for enhancing visual content, with significant applications in creative industries. The advent of GANs (Goodfellow et al. 2020; Karras, Laine, and Aila 2019) and diffusion models (Saharia et al. 2022; Rombach et al. 2022; Gu et al. 2022) has greatly improved image editing, achieving high levels of realism and precision. Conditional editing, which has gained considerable interest, enables users to image personalization. Commonly, it can be classified into text-conditioned (Bar-Tal et al. 2022; Bermano et al. 2022; Kim, Kwon, and Ye 2022a) and image-conditioned editing (Yang et al. 2023) based on the type of control input. These methods fine-tune generative models for specific conditions, producing high-fidelity images. Meanwhile, the large-scale brain activity datasets recording functional magnetic resonance imaging (fMRI) or electroencephalogram (EEG) have significantly advanced the field of brain decoding (Scotti et al. 2023; Chen et al. 2023). It decodes the implicit brain signals into explicit representations like images. However, brain signal, as a modal of information, has been seldom utilized for image editing. Previous works primarily focus on facial editing with only a few brain attributes as conditions (Davis, de la Torre-Ortiz, and Ruotsalo 2022), hindering the widespread use. Therefore, we raise the question that *how about brain signals as prompts enable the editing of more generic images?*

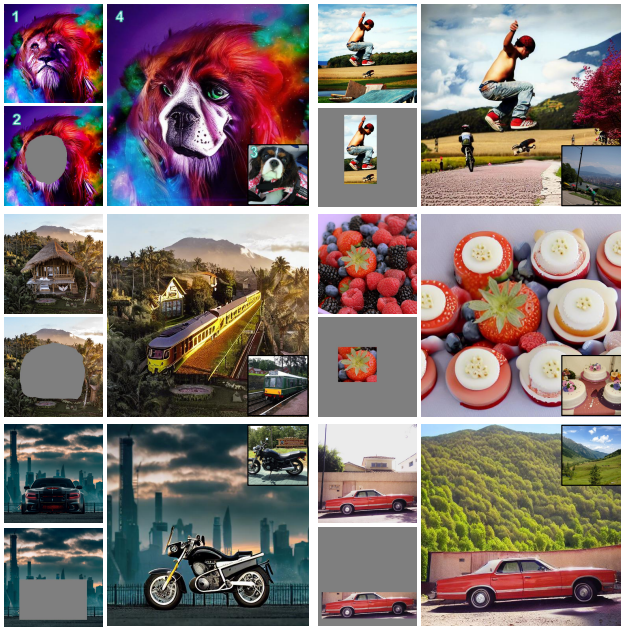

Figure 1: MindPainter performs high-quality, consistent editing on brain-conditioned painting tasks. Here we provide the fMRI of visual stimuli to paint the results, where annotations 1, 2, 3, and 4 represent the source, masked image, visual stimuli, and generated image, respectively.

Although recent advancements in text and image-

*These authors contributed equally.
†Corresponding Author

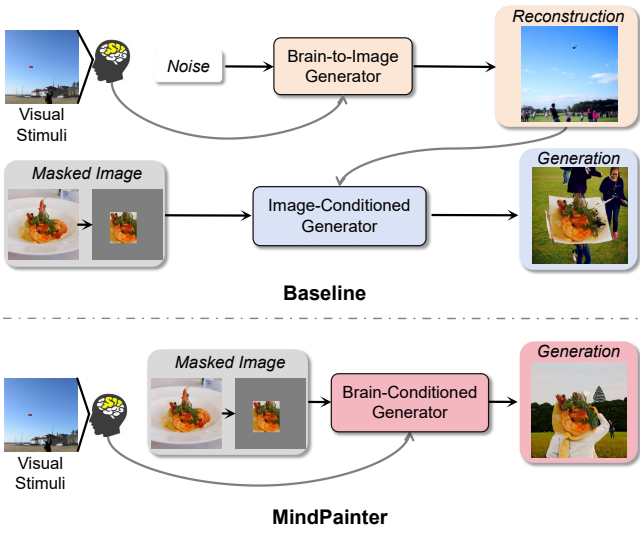

Figure 2: Inference comparison. The baseline represents the naive combinations of fMRI-to-image reconstruction and image-conditioned editing. Our MindPainter generates brain-conditioned results in a straightforward manner.

conditioned editing achieve realistic images, these conditions are indirect and complex when expressing the real users' mind. Brain signals provide a more direct and accurate recording of brain activities, allowing for personalized and human-centered image editing. It simplifies the process to mere "thinking" and holds significant potential for brain-computer interfaces. Currently, fMRI of visual stimuli is the most abundant and widely researched data. These signals are recorded when humans view images of natural scenes, which contain rich semantic information. Thereby, we hope to explore the visual brain signals for region-based natural image editing (*image painting*), making human thoughts a prospective prompt.

One intuitive method is to use the combination of existing methods, fMRI-to-image reconstruction (Scotti et al. 2023), and image-conditioned editing (Yang et al. 2023). It first reconstructs fMRI to images of visual stimuli and then uses these images as prompts for editing. As shown in Figure 2, this two-step process suffers from inefficient inference, biased brain interpretation, and unnatural editing results. We summarize the reasons for this failure which are: (1) naive combination, (2) gap across modalities, and (3) limited capacity of existing methods.

To achieve efficient and naturally edited outcomes, we aim to address the following three aspects in our design: (1) integrating the brain interpretation and image painting within a single pipeline for simplified model inference; (2) ensuring that the unified pipeline naturally merges both modalities for consistent editing; (3) solving the lack of editing labels in the brain-conditioned image painting task.

Motivated by the above analysis, we propose MindPainter, a novel cross-modal self-supervised learning for efficient and natural brain-conditioned image painting. MindPainter is trained for reconstruction from masked images via directly leveraging the information of brain signals. The brain signals are simulated by our proposed Pseudo Brain Generator. Thus, we integrate semantics from both brain and image modalities into a unified process, laying a crucial foundation for generating results that are blended efficiently and naturally. To further achieve a seamless and coherent integration of the two modalities in one generated image, we propose Brain Adapter to map brain signals into latent space, which shares common representation with image modal. Besides, Multi-Mask Generation Policy is designed to enhance the model's generalization. This enables our method to better adapt to different types of masked regions for image painting, including inpainting, outpainting and even reconstruction.

Our contributions are: (1) We propose a novel cross-modal self-supervised learning for brain-conditioned image painting. It integrates semantic information from both brain and image modalities into a unified process, enabling efficient and natural editing. (2) To accommodate this task, we design unique modules, the Pseudo Brain Generator and the Brain Adapter. Besides, we introduce the Multi-Mask Generation Policy to further enhance the model performance. (3) Our method is capable of efficiently generating high-quality editing results across various painting scenarios. To the best of our knowledge, MindPainter is the first work that achieves brain-conditioned editing on natural images.

Related Work

Driven by GANs (Ren et al. 2019; Zeng et al. 2019; Patashnik et al. 2021a,b) and diffusion (Avrahami, Lischinski, and Fried 2022; Hertz et al. 2023; Kawar et al. 2023; Nichol et al. 2022; Kim, Kwon, and Ye 2022b; Ruiz et al. 2023), image painting (Ballester et al. 2001; Bertalmío et al. 2003; Yang et al. 2023) has advanced to produce realistic and contextually consistent results. Conditional painting, which includes text-conditioned (Nichol et al. 2022; Kim, Kwon, and Ye 2022b; Suvorov et al. 2022a; Yu et al. 2018; Su et al. 2023) and image-conditioned methods (Yang et al. 2023), has become popular, enabling users to edit images with personalized controls. These methods fine-tune the diffusion models for specific conditions to synthesize high-quality images. Meanwhile, the field of brain decoding (Cox and Savoy 2003; Horikawa and Kamitani 2015; Schoenmakers et al. 2013; VanRullen and Reddy 2018; Gu et al. 2023; Quan et al. 2024; Wang et al. 2024) has seen substantial advancements in recent years. Efforts on large-scale visual brain signal datasets (Allen 2022; Chang 2019) further enhance the exploration of human brain signals from visual stimuli. Combined with diffusion models (Rombach et al. 2022), it has promoted visual brain signal decoding via conditioning on brain signals (Takagi and Nishimoto 2023), incorporating contrastive learning (Scotti et al. 2023; Chen, Qing, and Zhou 2023) and masked modeling (Chen et al. 2023). Furthermore, very few studies (Davis, de la Torre-Ortiz, and Ruotsalo 2022) have explored the use of brain signals as supervision for image editing. However, their works have been limited to only a few attributed conditions of brain signals for GAN-based facial editing. This paper pioneers the use of brain signals of various visual stimuli as the condition for

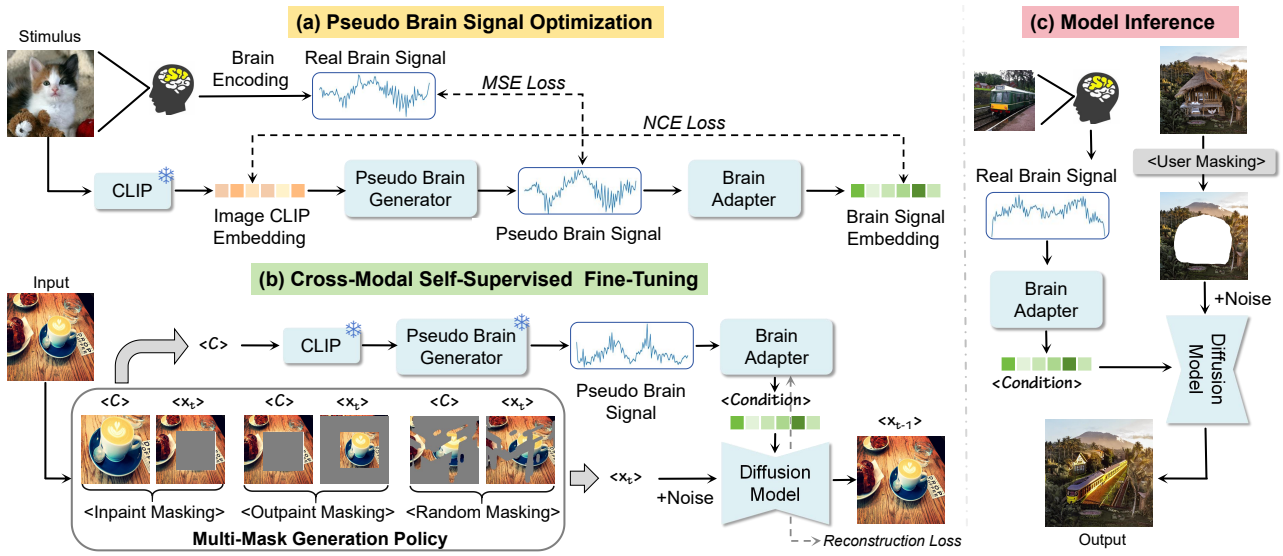

Figure 3: Training and inference overview of MindPainter. (a) We train the Pseudo Brain Generator (PBG) to map CLIP-embedded image to pseudo-brain signal, and optimize the Brain Adapter (BA) to embed brain signal. (b) We finetune the diffusion model with conditioning BA to reconstruct image across brain-image modalities. We randomly select one of the masking policies for each input. (c) We feed the masked image with brain condition to directly generate the editing result.

natural image painting. It efficiently achieves high-quality editing results.

Method

In the brain-conditioned image painting task, the goal is to integrate brain signal conditions into a source image, ensuring the resulting image appears plausible and photo-realistic. Let $\mathbf{x}_s \in R^{H\times W\times 3}$ represent the source image, where H and W denote the height and width, respectively. The region to be edited is specified by a binary mask $\mathbf{m} \in \{0,1\}^{H\times W}$, where a value of 1 marks the editable positions in $\mathbf{x}_s$. With the condition of brain signal $\mathbf{x}_b$, the objective is to create a new image $\mathbf{y}$ from $(\mathbf{x}_s, \mathbf{x}_b, \mathbf{m})$, where the area of $\mathbf{m} = 1$ should seamlessly integrate the semantics of brain signal $\mathbf{x}_b$.

Why existing mature methods fail in the task of brain-conditioned painting on natural images? Intuitively, we can combine two state-of-the-art techniques, which are MindEye (Scotti et al. 2023) and Paint by Example (Yang et al. 2023) to achieve this task. MindEye first reconstructs the implicit fMRI of visual stimuli to an explicit image. Then, Paint by Example uses the reconstructed image as the prompt for image painting. Although the two-step approach yields decent results, it presents several issues due to the naive combination. Firstly, simply stacking two generative processes, which both require pre-trained diffusion models, leads to high computational memory and inference time. The inefficiency makes image editing impractical. Secondly, the two steps increase the likelihood of biases in interpreting the brain signals, making the outcome of the second painting step heavily reliant on the first decoding accuracy. If the brain decoding is biased, the subsequent image painting will inevitably be flawed. Thirdly, the leverage of existing methods suffers from the limited model capacity. It can not well generalize across different painting scenarios, resulting in poor stylistic consistency and unnatural editing results.

Our Method

Cross-Modal Self-Supervised Pipeline The training and inference overview of MindPainter is shown in Figure 3. In this paper, we aim to integrate the interpretation of brain signals with image painting within a single pipeline, achieving efficient and natural editing. Considering the lack of editing labels, we introduce self-supervised learning (Balestriero et al. 2023) for our cross-modal pipeline. It reconstructs the source from the masked image by directly utilizing the information from a brain signal. Specifically, given a source image $\mathbf{x}_s$ and its corresponding bounding box, we utilize the bounding box as a binary mask $\mathbf{m}$ to obtain the masked image $\overline{\mathbf{m}} \odot \mathbf{x}_s$ and complementary image $\mathbf{m} \odot \mathbf{x}_s$. To simulate the visual brain signal from the complementary image, we propose the **Pseudo Brain Generator** to address this problem. Consequently, we can obtain the pseudo-brain condition $\mathbf{x}_b$ via $\mathbf{x}_b = \mathcal{G}_\eta(\mathbf{m} \odot \mathbf{x}_s)$, where $\mathcal{G}_\eta$ is parameterized with η as the Pseudo Brain Generator. Further, we propose the **Brain Adapter** $\mathcal{A}_\sigma$ parameterized with σ to embed the semantics of pseudo-brain signals to reduce the gap between image and brain modalities in the implicit space. Finally, the training objective is to reconstruct the source image $\mathbf{x}_s$ from the masked image and pseudo-brain condition. The proposed cross-modal self-supervised pipeline integrates semantically information from two modalities into a unified process, laying a crucial foundation for generating results that are blended efficiently and naturally.

Pseudo Brain Signal Optimization To simulate the brain signal for complementary image, we propose the Pseudo Brain Generator (PBG) to produce the simulated brain sig-

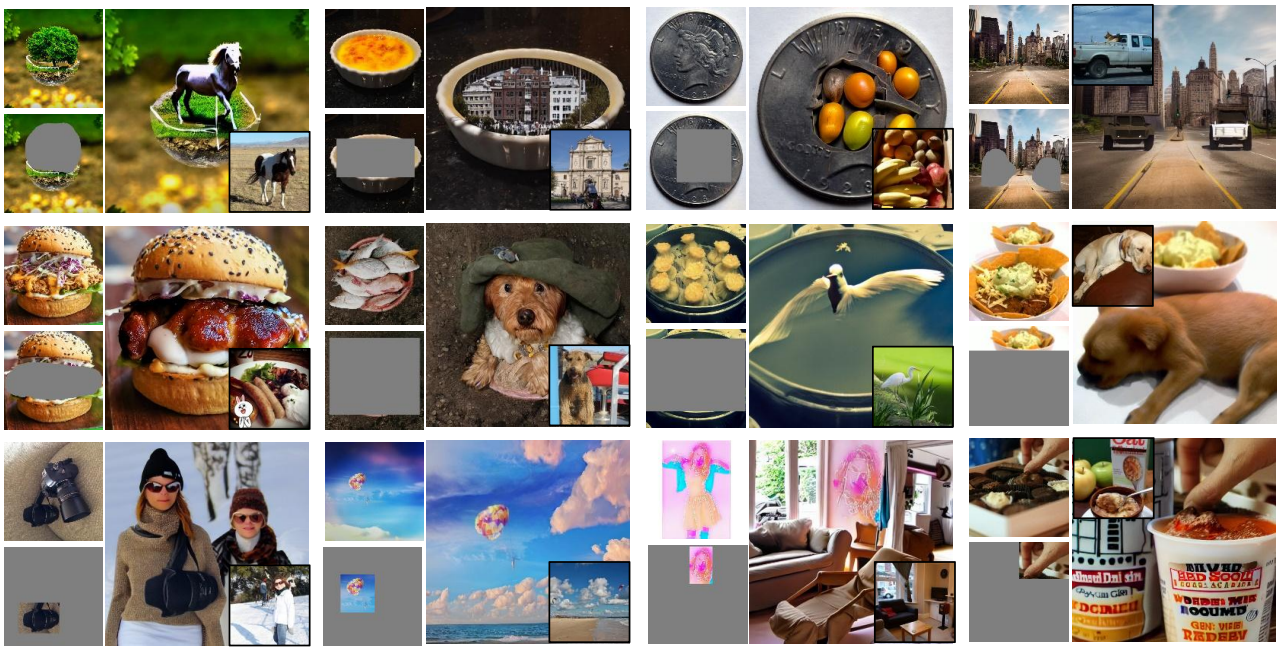

Figure 4: Qualitative results. MindPainter generates high-quality editing results with various masked source images and visual stimuli conditions (thumbnail). It can be applied to various painting scenarios, including inpainting (the first and second lines) and outpainting (the third line).

nal. Because of the strong regression ability of Multilayer Perception (MLP), we adopt the MLP backbone for PBG, which consists of several linear layers with residual blocks. We leverage the NSD dataset for the supervised training of PBG, which comprises paired data (B, I), where B denotes the fMRI brain signal of a human viewing the natural image I. The brain data $B \in R^{N\times V}$ is a vector that contains V voxels in the batch-size of N. The image $I \in R^{N\times 256\times 256}$ is fed into the pretrained CLIP ViT/L-14 (Radford et al. 2021) to obtain the intermediate feature Z of size $N \times 1024$. The feature is then passed through the PBG to regress the brain signal B^*, which is supervised by ground truth B using the MSE loss. The MSE loss can be formulated as:

$$\begin{aligned} Z_i &= CLIP_{Image}(I_i), \\ \mathcal{L}_{PBG} &= \sum_{i=1}^{N} L_{MSE}(\mathcal{G}(Z_i, \eta), B_i), \end{aligned} \tag{1}$$

where I_i, Z_i and B_i represents the i_{th} image, feature and real brain signal, respectively. L_{MSE} is the MSE loss function. Note that the CLIP model remains frozen.

Further, we feed brain signals into the designed Brain Adapter $\mathcal{A}_\sigma$ to get the brain embedding. The Brain Adapter (BA) is also an MLP backbone consisting of several linear layers with residual blocks. The unified architecture of PBG and BA is a symmetric decoding-encoding structure, providing convenient model optimization. To achieve the unified embedding space, we utilize contrastive learning to encode the brain modal into the embedding space of the pre-trained CLIP-image modal. To facilitate the optimization, we feed both real and simulated brain signals B and B^* into the BA to obtain the embedding $E \in R^{N\times 1024}$ and $E^* \in R^{N\times 1024}$. The similarity matrix S is computed as follows:

$$S = Z \cdot E^\top,$$

where S_{ij} represents the similarity between image Z_i and real brain embedding E_j. Similarly, for Z and E^*, the similarity matrix S^* is computed as follows:

$$S^* = Z \cdot E^{*\top}.$$

Thus, the CLIP contrastive loss is performed on the S and S^*, respectively:

$$\mathcal{L}_{\text{real\_contra}} = -\frac{1}{N}\sum_{i=1}^{N}\left[\log\frac{\exp(S_{ii}/\tau)}{\sum_{j=1}^{N}\exp(S_{ij}/\tau)} + \log\frac{\exp(S_{ii}/\tau)}{\sum_{j=1}^{N}\exp(S_{ji}/\tau)}\right], \tag{2}$$

$$\mathcal{L}_{\text{simul\_contra}} = -\frac{1}{N}\sum_{i=1}^{N}\left[\log\frac{\exp(S^*_{ii}/\tau)}{\sum_{j=1}^{N}\exp(S^*_{ij}/\tau)} + \log\frac{\exp(S^*_{ii}/\tau)}{\sum_{j=1}^{N}\exp(S^*_{ji}/\tau)}\right]. \tag{3}$$

The overall CLIP contrastive loss can be formulated as:

$$\mathcal{L}_{\text{contra}} = \mathcal{L}_{\text{real\_contra}} + \mathcal{L}_{\text{simul\_contra}}. \tag{4}$$

For mutually facilitating the training of the PBG and BA, here we merge the training of $\mathcal{G}_\eta$ and $\mathcal{A}_\sigma$ for joint optimization. The Pseudo Brain Signal Optimization provides the foundation for accurate exaction of pseudo-brain signals and corresponding semantics.

Diffusion Prior Diffusion models have made significant strides in producing exceptionally high-quality images and have been effectively utilized in numerous text and image-conditioned editing tasks. In this context, we employ the dif-

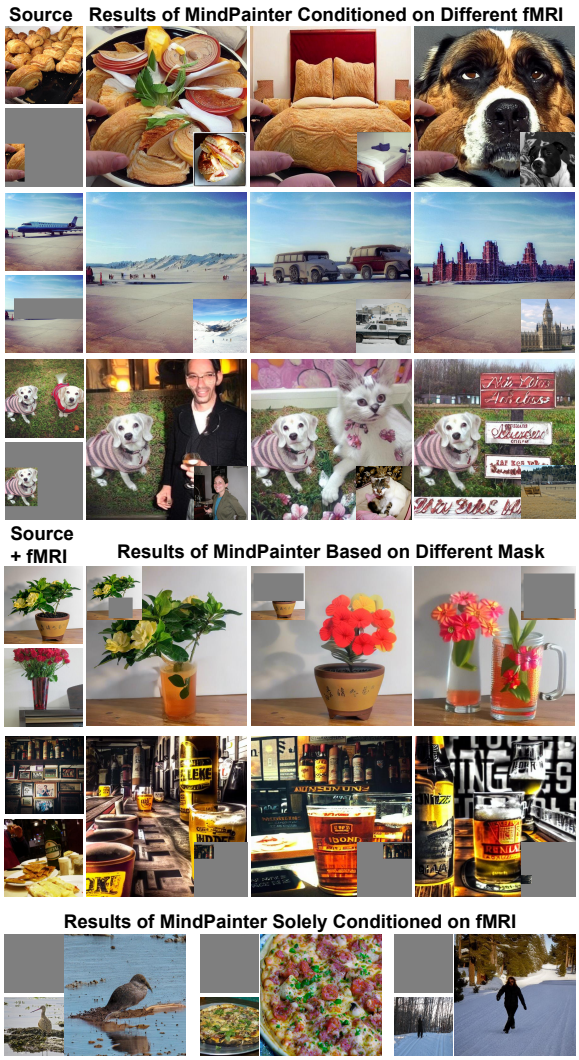

Figure 5: More painting cases.

fusion model pre-trained on image-driven editing for initialization (Yang et al. 2023), serving as a robust image prior.

Cross-Modal Self-Supervised Fine-Tuning To achieve a seamless and coherent integration of brain and image modalities in the generated images, it is imperative to enable mutual understanding and fusion of the two types of information within the generative model. The cross-modal self-supervised fine-tuning incorporates the BA $\mathcal{A}_\sigma$ as a conditioning module into the diffusion model for joint fine-tuning. It reconstructs the source image in latent diffusion by embedding the prompt from the pseudo-brain signal. This process further enhances the mapping ability of BA and promotes the coherent consolidation of two modalities. To ensure steady convergence, we fix the PBG in conditional diffusion fine-tuning. Following the conditional probability optimization of the diffusion model, the training objective function can be formulated as:

$$\mathcal{L}_t^{\text{cond}} = E_{x,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(x_t, t, \mathcal{A}(B^*, \sigma))\|_2^2\right]. \quad (5)$$

where $t = 1, \ldots, T$ and x_t is obtained by corrupting the masked image x_{t-1} with Gaussian noise. $\epsilon_\theta(x_t, t)$ is a set of denoising functions that are usually implemented as UNets, referring to (Rombach et al. 2022) for more detailed descriptions of stable diffusion.

Multi-Mask Generation Policy In fine-tuning, we introduce Multi-Mask Generation Policy to increase training complexity, thereby enhancing the model's robustness and generalization. This enables our model to better adapt to different masked regions for image painting. We employ three specific masking strategies: inpaint, outpaint, and random masking. Specifically, based on the bounding box provided by $\mathbf{x}_s$, inpaint masking treats the image patch within the bounding box as the condition, outpaint masking uses the patch outside the bounding box. Drawing on the method from LAMA (Suvorov et al. 2022b), random masking is applied to $\mathbf{x}_s$ to obtain masks of random shapes, positions, and quantities. Inspired from (Yang et al. 2023), we apply mask shape augmentation to all the aforementioned masks for better adaptation to user-masked painting scenarios. Note that the PBG, which is trained for mapping the CLIP-embedded image into the brain modal, is capable of dealing with different masking image patches. During training, we utilize a probabilistic selection to choose a masking strategy for the current training image. The manually-set probabilities for the inpaint, outpaint, and random masking strategy are 0.5, 0.3, and 0.2, respectively.

Inference Stage We directly use the real brain signal of visual stimulus as the condition and a user-masked image as the diffusion input. Specifically, the brain signal is first fed into the BA for embedding and then injected into the diffusion model via cross-attention. Finally, we can efficiently paint the masked image with the brain condition in a straightforward manner, in which the edited image is stylistically consistent, naturally coherent, and logically sound.

Experiment

Dataset and Implementation

The NSD dataset (Allen 2022) is a large-scale fMRI dataset collecting the brain responses of human visual perception when viewing natural scenes from MS-COCO (Lin et al. 2014). We utilize it for Pseudo Brain Signal Optimization. Here we develop subject-specific models for each of the four subjects in NSD. We present the results of subject 1 in this paper. In fine-tuning, we select 10,000 images from OpenImages dataset in proportion to the original distribution of 600 categories. For inference illustration, the source images are collected from OpenImages and free-to-use images from the Bing website, and fMRI from the NSD test set is applied as the condition. In the user study, we randomly pair 100 source images from OpenImages with 100 fMRI from NSD as our test benchmark.

Qualitative Illustration

We apply MindPainter to various painting scenarios, including inpainting and outpainting tasks with arbitrary masks created by users. As shown in Figure 4, MindPainter enables the seamless integration of implicit brain signal semantics

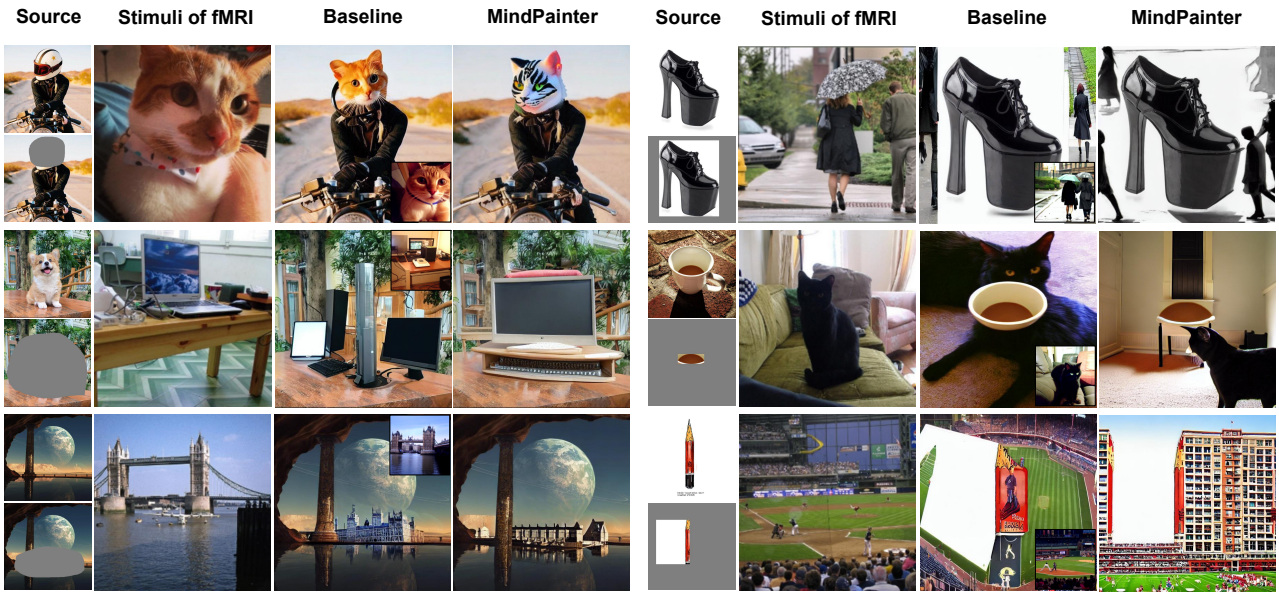

Figure 6: Method comparison. We compare MindPainter with the baseline in inpainting (left) and outpainting (right) tasks. Overall, MindPainter generates editing results that are stylistically consistent and naturally coherent.

Method	Quality↑	Consistency↑	Alignment↑
Baseline	$\mathbf{2.49 \pm 0.62}$	1.98 ± 0.82	2.11 ± 0.81
MindPainter	2.47 ± 0.62	$\mathbf{2.31 \pm 0.75}$	$\mathbf{2.36 \pm 0.76}$
Diffusion Prior	1.27 ± 0.53	1.37 ± 0.66	1.15 ± 0.47
wo PBG	1.49 ± 0.55	1.32 ± 0.56	1.39 ± 0.54
+ inpaint masking	2.17 ± 0.80	2.08 ± 0.82	1.54 ± 0.54
+ in & out masking	2.29 ± 0.71	2.17 ± 0.67	2.02 ± 0.73

Table 1: User study of quantitative comparison. Users are requested to score on the metrics of image quality, semantic consistency, and condition alignment from 1 to 3 (1 is the worst, 3 is the best). MindPainter can achieve high-quality editing results with superior consistency and alignment.

into natural image edits. Furthermore, we apply our method to two scenarios commonly encountered by real users: the same source image with different brain conditions, and the same source image and brain condition with different masks. As shown in Figure 5, our method demonstrates excellent robustness and generalization capabilities. Regardless of the variations in brain conditions and masks, our method consistently generates results that are natural and logically coherent. Moreover, we tested the effectiveness of our method under extreme situations: when the entire source image is masked, the image editing relies solely on the brain condition. Even in such cases, our method can generate images that accurately reproduce the semantics of the brain signals.

Comparisons

Qualitative Analysis We adopt two state-of-the-art methods, MindEye and Paint by Example as the brain decoder and image-based editing strategy to get the combination as the baseline in our paper. In Figure 6, we qualitatively compare MindPainter with the baseline. The results show that while the baseline can generate edited images with details, the naive concatenation of the two methods increases the likelihood of semantic deviation in the brain signal. Moreover, constrained by the editing capability, the baseline generates images that lack natural consistency. This is particularly evident in outpainting tasks. In contrast, MindPainter directly reflects the semantic information of brain signals, seamlessly integrating the condition with the source image in both inpainting and outpainting tasks.

User Study To present the quantitative analysis of the painting results, we conduct the human perceptual evaluation study on 20 participants, who are divided into 5 groups. Each group is evaluated on 20 pairs of comparison, including our results and the baseline results. Participants are asked to score on three perspectives independently: the generated image quality, the alignment to the semantics of the brain signal, and the consistency. In total, we collected 1200 answers, whose results are summarized in Table 1. Notably, evaluators exhibit a preference for our method, especially on the consistency and alignment metrics.

Ablation Study

As shown in Figure 7 and Table 1, we conduct ablations with user study on three key technologies introduced in our method to compare their effectiveness: diffusion prior, PBG, and Multi-Mask Generation Policy. (1) The gap between the two modalities results in poor detail and quality. (2) The results indicate that the edited images appear highly unnatural, exhibiting noticeable editing boundaries. We believe the model fails to learn the relationship between the condition and the masked image. It proves the simulated brain signal from the complementary image is crucial. (3) With more so-

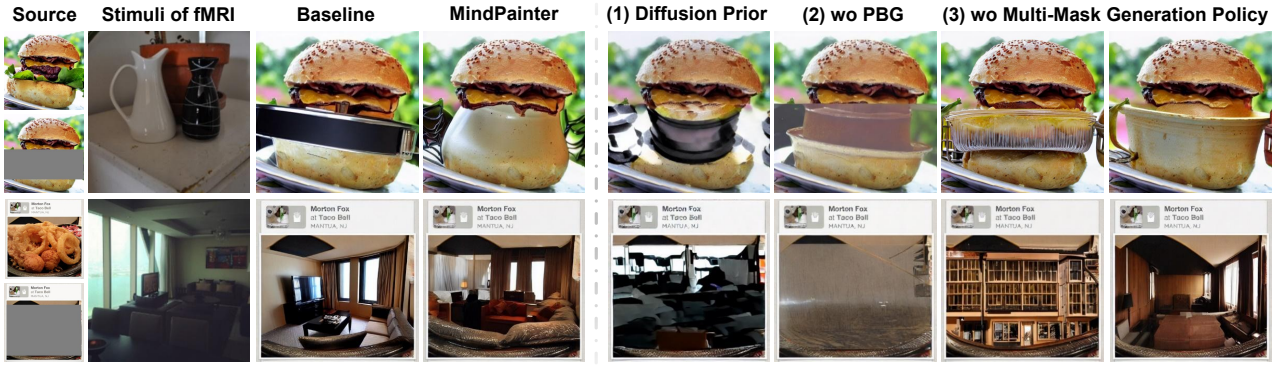

Figure 7: Qualitative ablations. We demonstrate the crucial role of different techniques in MindPainter. (1) We directly utilize diffusion prior with brain conditions for editing. (2) Instead of generating pseudo-brain signals from *complementary* images, the simulated fMRI of the *entire* source image is used for training. (3) We stack masking policies one by one: only inpaint masking (left) and added with outpaint masking (right).

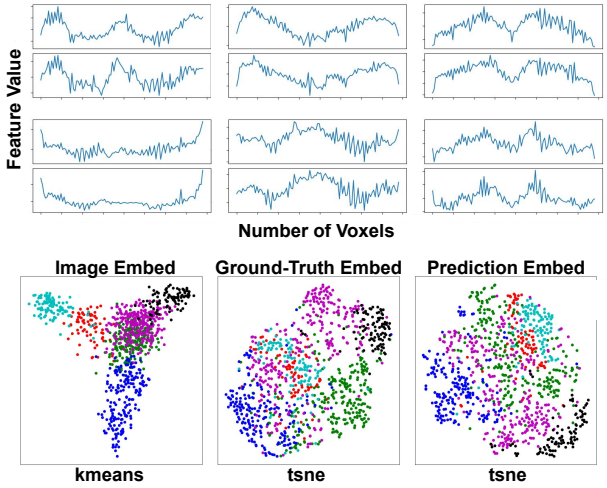

Figure 8: Visualization of predicted pseudo-fMRI with ground truth, where the first line is the ground truth and the second is the predicted voxels. The horizontal axis denotes the number of voxels, ranging from 0∼16000 and the vertical axis is the feature value of voxels, ranging from -1∼1. The clustering of semantic embeddings is illustrated below.

phisticated masking, models generate more detailed and natural images across different painting scenarios.

Effectiveness of Designed Modules

In Figure 8, we evaluate the effectiveness of the designed modules, PBG and BA. (1) We visualize the generated pseudo-brain signals with the ground truth. The illustration represents the values of voxels in a fMRI signal. The pseudo-fMRI almost matches the value of the ground truth. The results indicate that the PBG is capable of mapping CLIP-embedded images to the brain modal. (2) To further evaluate the accuracy of both real and pseudo fMRI semantics, we present the experiment on semantic clustering. Following the clustering setting of the NSD dataset, the images of

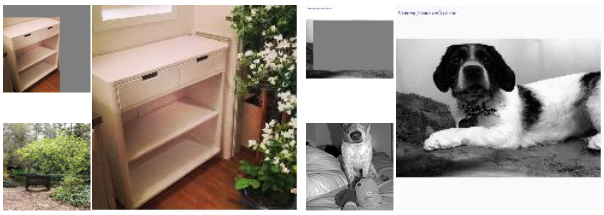

Figure 9: Qualitative effectiveness of designed modules. We utilize the PBG to generate pseudo-fMRI from a reference image (left below) and then apply the pseudo-fMRI for image editing.

visual stimuli are classified into six labels, which are people, animals, inanimate, people+animals, people+inanimate, and inanimate+animals. We obtain the CLIP-embedded images and perform k-means (Ikotun et al. 2023) supervised clustering on six labels. Then the corresponding real and pseudo-fMRI are embedded via the BA and clustered via unsupervised tsne (Linderman et al. 2017). It is observed that both the embeddings of real and pseudo-fMRI are well-clustered. It demonstrates that the BA accurately captures the CLIP-like features of fMRI, meanwhile, the pseudo-fMRI by PBG not only simulates the voxel value but also preserves the semantics. (3) Figure 9 qualitatively demonstrates the effectiveness of our designed modules.

Conclusion

In this paper, we propose a novel cross-modal self-supervised learning pipeline MindPainter, using fMRI of visual stimuli as prompts for natural image painting. By reconstructing from masked images with pseudo-brain signals simulated by the PBG and ensuring accurate brain signal interpretation with the BA, we achieve efficient and seamless painting in various scenarios. MindPainter is the first efficient brain-conditioned painting of natural images, enhancing brain-computer interaction for creative AI.

Acknowledgments

This work was supported by the Institute for Interdisciplinary Information Core Technology (IIISCT), Xi'an, China.

References

Allen, S.-Y. G. W. Y. e. a., E.J. 2022. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. 116–126.

Avrahami, O.; Lischinski, D.; and Fried, O. 2022. Blended Diffusion for Text-driven Editing of Natural Images. In *CVPR*, 18187–18197. IEEE.

Balestriero, R.; Ibrahim, M.; Sobal, V.; Morcos, A.; Shekhar, S.; Goldstein, T.; Bordes, F.; Bardes, A.; Mialon, G.; Tian, Y.; Schwarzschild, A.; Wilson, A. G.; Geiping, J.; Garrido, Q.; Fernandez, P.; Bar, A.; Pirsiavash, H.; LeCun, Y.; and Goldblum, M. 2023. A Cookbook of Self-Supervised Learning. *CoRR*, abs/2304.12210.

Ballester, C.; Bertalmío, M.; Caselles, V.; Sapiro, G.; and Verdera, J. 2001. Filling-in by joint interpolation of vector fields and gray levels. *IEEE Trans. Image Process.*, 10(8): 1200–1211.

Bar-Tal, O.; Ofri-Amar, D.; Fridman, R.; Kasten, Y.; and Dekel, T. 2022. Text2LIVE: Text-Driven Layered Image and Video Editing. In *ECCV (15)*, volume 13675 of *Lecture Notes in Computer Science*, 707–723. Springer.

Bermano, A. H.; Gal, R.; Alaluf, Y.; Mokady, R.; Nitzan, Y.; Tov, O.; Patashnik, O.; and Cohen-Or, D. 2022. State-of-the-Art in the Architecture, Methods and Applications of StyleGAN. *Comput. Graph. Forum*, 41(2): 591–611.

Bertalmío, M.; Vese, L. A.; Sapiro, G.; and Osher, S. J. 2003. Simultaneous Structure and Texture Image Inpainting. In *CVPR (2)*, 707–712. IEEE Computer Society.

Chang, P.-J. M. A. e. a., N. 2019. BOLD5000, a public fMRI dataset while viewing 5000 visual images.

Chen, Z.; Qing, J.; Xiang, T.; Yue, W. L.; and Zhou, J. H. 2023. Seeing Beyond the Brain: Conditional Diffusion Model with Sparse Masked Modeling for Vision Decoding. In *CVPR*, 22710–22720. IEEE.

Chen, Z.; Qing, J.; and Zhou, J. H. 2023. Cinematic Mindscapes: High-quality Video Reconstruction from Brain Activity. *NeurIPS*.

Cox, D. D.; and Savoy, R. L. 2003. Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage*, 19(2): 261–270.

Davis, K. M.; de la Torre-Ortiz, C.; and Ruotsalo, T. 2022. Brain-Supervised Image Editing. In *CVPR*, 18459–18468. IEEE.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A. C.; and Bengio, Y. 2020. Generative adversarial networks. *Commun. ACM*, 63(11): 139–144.

Gu, S.; Chen, D.; Bao, J.; Wen, F.; Zhang, B.; Chen, D.; Yuan, L.; and Guo, B. 2022. Vector Quantized Diffusion Model for Text-to-Image Synthesis. In *CVPR*, 10686–10696. IEEE.

Gu, Z.; Jamison, K.; Kuceyeski, A.; and Sabuncu, M. R. 2023. Decoding natural image stimuli from fMRI data with a surface-based convolutional network. In *MIDL*, volume 227 of *Proceedings of Machine Learning Research*, 107–118. PMLR.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2023. Prompt-to-Prompt Image Editing with Cross-Attention Control. In *ICLR*. OpenReview.net.

Horikawa, T.; and Kamitani, Y. 2015. Generic decoding of seen and imagined objects using hierarchical visual features. *CoRR*, abs/1510.06479.

Ikotun, A. M.; Ezugwu, A. E.; Abualigah, L.; Abuhaija, B.; and Jia, H. 2023. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Inf. Sci.*, 622: 178–210.

Karras, T.; Laine, S.; and Aila, T. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *CVPR*, 4401–4410. Computer Vision Foundation / IEEE.

Kawar, B.; Zada, S.; Lang, O.; Tov, O.; Chang, H.; Dekel, T.; Mosseri, I.; and Irani, M. 2023. Imagic: Text-Based Real Image Editing with Diffusion Models. In *CVPR*, 6007–6017. IEEE.

Kim, G.; Kwon, T.; and Ye, J. C. 2022a. DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation. In *CVPR*, 2416–2425. IEEE.

Kim, G.; Kwon, T.; and Ye, J. C. 2022b. DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation. In *CVPR*, 2416–2425. IEEE.

Lin, T.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *ECCV (5)*, volume 8693 of *Lecture Notes in Computer Science*, 740–755. Springer.

Linderman, G. C.; Rachh, M.; Hoskins, J. G.; Steinerberger, S.; and Kluger, Y. 2017. Efficient Algorithms for t-distributed Stochastic Neighborhood Embedding. *CoRR*, abs/1712.09005.

Nichol, A. Q.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, 16784–16804. PMLR.

Oh, B. M.; Chen, M.; Dorsey, J.; and Durand, F. 2001. Image-based modeling and photo editing. In *SIGGRAPH*, 433–442. ACM.

Patashnik, O.; Wu, Z.; Shechtman, E.; Cohen-Or, D.; and Lischinski, D. 2021a. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. In *ICCV*, 2065–2074. IEEE.

Patashnik, O.; Wu, Z.; Shechtman, E.; Cohen-Or, D.; and Lischinski, D. 2021b. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. In *ICCV*, 2065–2074. IEEE.

Quan, R.; Wang, W.; Tian, Z.; Ma, F.; and Yang, Y. 2024. Psychometry: An Omnifit Model for Image Reconstruction from Human Brain Activity. *CoRR*, abs/2403.20022.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.

Ren, Y.; Yu, X.; Zhang, R.; Li, T. H.; Liu, S.; and Li, G. 2019. StructureFlow: Image Inpainting via Structure-Aware Appearance Flow. In *ICCV*, 181–190. IEEE.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*, 10674–10685. IEEE.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *CVPR*, 22500–22510. IEEE.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, S. K. S.; Lopes, R. G.; Ayan, B. K.; Salimans, T.; Ho, J.; Fleet, D. J.; and Norouzi, M. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *NeurIPS*.

Schoenmakers, S.; Barth, M.; Heskes, T.; and van Gerven, M. 2013. Linear reconstruction of perceived images from human brain activity. *NeuroImage*, 83: 951–961.

Scotti, P. S.; Banerjee, A.; Goode, J.; Shabalin, S.; Nguyen, A.; Cohen, E.; Dempster, A. J.; Verlinde, N.; Yundler, E.; Weisberg, D.; Norman, K. A.; and Abraham, T. M. 2023. Reconstructing the Mind's Eye: fMRI-to-Image with Contrastive Learning and Diffusion Priors. In *NeurIPS*.

Su, X.; Song, J.; Meng, C.; and Ermon, S. 2023. Dual Diffusion Implicit Bridges for Image-to-Image Translation. In *ICLR*. OpenReview.net.

Suvorov, R.; Logacheva, E.; Mashikhin, A.; Remizova, A.; Ashukha, A.; Silvestrov, A.; Kong, N.; Goka, H.; Park, K.; and Lempitsky, V. 2022a. Resolution-robust Large Mask Inpainting with Fourier Convolutions. In *WACV*, 3172–3182. IEEE.

Suvorov, R.; Logacheva, E.; Mashikhin, A.; Remizova, A.; Ashukha, A.; Silvestrov, A.; Kong, N.; Goka, H.; Park, K.; and Lempitsky, V. 2022b. Resolution-robust Large Mask Inpainting with Fourier Convolutions. In *WACV*, 3172–3182. IEEE.

Takagi, Y.; and Nishimoto, S. 2023. High-resolution image reconstruction with latent diffusion models from human brain activity. In *CVPR*, 14453–14463. IEEE.

VanRullen, R.; and Reddy, L. 2018. Reconstructing Faces from fMRI Patterns using Deep Generative Neural Networks. *CoRR*, abs/1810.03856.

Wang, S.; Liu, S.; Tan, Z.; and Wang, X. 2024. MindBridge: A Cross-Subject Brain Decoding Framework. *CoRR*, abs/2404.07850.

Yang, B.; Gu, S.; Zhang, B.; Zhang, T.; Chen, X.; Sun, X.; Chen, D.; and Wen, F. 2023. Paint by Example: Exemplar-based Image Editing with Diffusion Models. In *CVPR*, 18381–18391. IEEE.

Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; and Huang, T. S. 2018. Generative Image Inpainting With Contextual Attention. In *CVPR*, 5505–5514. Computer Vision Foundation / IEEE Computer Society.

Zeng, Y.; Fu, J.; Chao, H.; and Guo, B. 2019. Learning Pyramid-Context Encoder Network for High-Quality Image Inpainting. In *CVPR*, 1486–1494. Computer Vision Foundation / IEEE.