Advancing Image Reconstruction: Evaluating BOLD and Beta Signals in fMRI-Based Brain Decoding

Anonymous CVPR submission

Paper ID

Abstract

Decoding visual experiences from brain activity presents a novel approach to comprehending how the brain perceives the world and to analyzing the link between our visual system and computer vision models. In particular, the Natural Scenes Dataset (NSD), which consists of fMRI Beta-Image pairs, is becoming a benchmark dataset. GLM-based fMRI Beta has been widely used in previous studies, but it has limitations in reflecting the complex and non-linear interactions between brain regions. We propose to investigate the performance of BOLD signals versus Beta values derived from GLM for fMRI-to-Image reconstruction, focusing on the MinD-Vis and MindEye models. Due to the limited training of our models, it is difficult to draw definitive conclusions about which type of fMRI data is more suitable for image reconstruction based on the current results. Future research should involve developing models that can account for interactions between brain regions, using large-scale fMRI datasets, and ensuring sufficient training time to revisit and thoroughly investigate the research question.

1. Introduction

Brain decoding, a field at the intersection of neuroscience and artificial intelligence, aims to interpret brain signals to extract meaningful information. By analyzing brain signals recorded while an individual watches a movie, researchers can potentially discern: 1) the specific scenes being viewed, 2) the content of the dialogue being heard, and 3) the emotions experienced at that moment. This challenging endeavor has seen significant progress due to recent advancements. Large-scale, publicly accessible brain imaging datasets, such as the UK Biobank, provide extensive data, including imaging data from thousands of individuals over several minutes of recording. Additionally, advancements in artificial intelligence and increased computational power have further facilitated brain decoding research.

In particular, brain decoding has shown promise in vision reconstruction. Deep learning models reconstructed images from fMRI signals [6], [20], [16]. This capability improves our understanding of visual processing in the brain and paves the way for applications in neuroprosthetics and brain-computer interfaces.

Building on this potential, recent research in brain decoding with non-invasive neuroimaging data, particularly fMRI-to-Image reconstruction, has shown significant progress. A widely used dataset in this domain is the NSD. The Natural Scenes Dataset (NSD) [2] comprises fMRI data collected from 8 participants who viewed a total of 73,000 RGB images. This dataset has been widely utilized to reconstruct perceived images from fMRI. However, the fMRI data from NSD and the models developed based on it have limitations for capturing the complex and nonlinear interactions between different brain regions essential for accurate brain vision decoding.

Previous studies [16], [20], [25] have utilized fMRI data derived not from the fMRI BOLD signal but from the Beta values obtained after applying a General Linear Model (GLM). The brain is a network system where different regions interact to perform functions. For instance, in facial recognition: 1) V1 extracts basic visual features such as edges, lines, and orientations from the visual information received from the retina; 2) V2 analyzes intermediate features by integrating patterns, textures, and spatial frequency information passed from V1; 3) V3 processes shape and motion information to help distinguish the contours and movements of faces; 4) V4 analyzes complex forms, including color information, to contribute to the formation of the overall image of a face; 5) Finally, the Fusiform Face Area (FFA) integrates all processed visual information to recognize faces. While the brain's visual processing involves both distinct regions and their interactions, the GLM assumes independence between voxels, failing to capture these interactions. Moreover, the brain's complex activities include nonlinear patterns [4], yet the Beta values from the GLM reflect only linear brain activity. Neuroscience research has demonstrated that applying nonlinear models, such as deep learning, to brain imaging data yields higher predictive performance compared to linear models [1].

Thus, we aim to investigate whether BOLD signals outperform Beta values for fMRI-to-Image reconstruction. Specifically, we will compare the performance of models using NSD's fMRI BOLD and fMRI Beta data in the context of Image Reconstruction, focusing on the MindEye and MinD-Vis models. MindEye has shown the best reconstruction performance using NSD data [2], while MinD-Vis, although not developed with NSD, was the first to apply large-scale pre-trained fMRI models in this domain. We anticipate that leveraging this pre-trained model will yield good reconstruction performance.

2. Related works

2.1. Brain Decoding

The evolution of brain decoding has seen significant advancements over recent decades, aiming to interpret brain activity patterns, particularly through non-invasive techniques like functional Magnetic Resonance Imaging (fMRI), to understand and reconstruct visual stimuli. Early works [10], [14] in this field primarily focused on classifying broad categories of visual stimuli using linear classifiers and basic machine learning techniques. [22] was the first to advance the field by applying convolutional neural networks (CNN) to extract image features and mapped fMRI signals to the CNN-based image features. These methods leveraged the hierarchical features extracted from pre-trained CNNs such as VGG [23] to map brain activity to visual categories.

Recent advancements have shifted towards reconstructing high-quality images directly from brain signals. [21] proposed an end-to-end deep learning framework that directly translates fMRI signals into images using a combination of autoencoders and generative adversarial networks (GANs). The advent of high-resolution image synthesis with Latent Diffusion Models [19] and multi-modal contrastive models like CLIP [18], along with the availability of extensive fMRI datasets [2], has significantly advanced research efforts in mapping fMRI signals into the CLIP embedding space. This technique facilitates latent diffusion models in image reconstruction, with various efforts through self-supervision [3], contrastive learning [20], and masked modeling [7]. A notable approach was proposed by [25], which utilizes Stable Diffusion to reconstruct images from fMRI by translating brain activities into text descriptions and subsequently generating corresponding images.

Despite these advancements, challenges such as the complexity of the visual cortex, remain significant hurdles. Studies have indicated that the complexity of representations within the visual cortex increases hierarchically [9]. In addition, [15] illustrated that leveraging information from various visual areas can enhance the performance of image reconstruction tasks. Consequently, simple decoding models without considering the non-linearity may be insufficient for accurate image reconstruction from brain activity.

2.2. Masked Brain Modeling

Masked Brain Modeling (MBM) represents a novel approach in the domain of self-supervised learning, specifically tailored for brain signal decoding. Inspired by the success of Masked Signal Modeling (MSM) in vision and language processing, MBM employs a similar strategy to learn effective representations from fMRI data. The core idea of MBM is to mask a portion of the input data and train a model to reconstruct the missing parts, thereby capturing the underlying structure and context of the data.

Chen [6] proposed the Sparse-Coded Masked Brain Modeling (SC-MBM) framework, which aligns with the biological principle of sparse coding observed in the visual cortex. In SC-MBM, fMRI data is divided into patches, and each patch is encoded into a high-dimensional vector space, creating an over-complete representation space. This approach not only enhances the capacity of the fMRI representations but also reflects the efficient coding strategies employed by the brain.

SC-MBM leverages large embedding-to-patch-size ratios and high mask ratios to exploit the spatial redundancy in fMRI data, enabling the model to learn rich and generalizable representations with minimal computational overhead. This technique has shown promise in generating more accurate and semantically meaningful reconstructions from fMRI data compared to conventional methods.

2.3. Latent Diffusion Model

Latent Diffusion Models (LDM) have emerged as powerful generative models capable of producing high-quality content by operating in the latent feature space. Unlike traditional diffusion models that work directly in the data space, LDMs compress images into a lower-dimensional latent space, which significantly reduces computational costs and improves image synthesis quality.

The LDM framework proposed by [19] incorporates a Vector Quantization (VQ) regularized autoencoder to compress images into latent features and a UNet-based denoising model with attention modules to perform the reverse diffusion process. This setup allows for flexible conditioning of image generation through cross-attention mechanisms, making LDMs highly suitable for tasks requiring conditional synthesis.

The effectiveness of LDMs in brain decoding tasks has been demonstrated through various researches. MindEye [20] is a notable example, which uses a novel fMRI-to-image approach that combines contrastive learning with a diffusion prior. This model comprises two parallel submodules specialized for retrieval and reconstruction, which enables mapping fMRI brain activity to a high-dimensional multimodal latent space, such as CLIP image space. [20] has shown state-of-the-art performance in both image reconstruction and retrieval tasks by leveraging advanced training techniques and large parameter models.

Another innovative approach is the BrainDiffuser [16]. The model employs a generative latent diffusion model for natural scene reconstruction from fMRI signals. BrainDiffuser effectively captures the semantic and structural aspects of the visual stimuli, producing high-quality reconstructions from brain activity data.

These advancements highlight the potential of LDMs in enhancing our understanding of the human visual system and improving brain-computer interfaces by achieving high-quality, semantically accurate image reconstruction from fMRI data.

3. Method

3.1. Dataset

We utilized the Natural Scenes Dataset (NSD) [2], a publicly available fMRI dataset that captures the brain responses of participants viewing natural scenes from the MS-COCO dataset [13]. The NSD [2] is a 7-Tesla fMRI dataset, comprising brain responses from several participants who each spent up to 40 hours in an MRI machine passively viewing images. These images, which are square-cropped and depict natural scenes, were sourced from the MS-COCO dataset [13]. Each of the 9,000-10,000 unique images was shown for three seconds, repeated three times across 30-40 scanning sessions, resulting in a total of 22,000-30,000 fMRI trials per participant.

We employed preprocessed, flattened fMRI voxels in 1.8-mm native volume space from the ”nsdgeneral” brain region, as defined by the NSD authors. This region includes approximately 16,000 voxels in the posterior cortex that are most responsive to the visual stimuli presented. The fMRI BOLD data, originally a 4D time series, was processed by extracting voxels from the ”nsdgeneral” region that were activated above a specific threshold and then flattening these into a 2D format. The fMRI beta data was obtained using GLMSingle [17], resulting in session-wise z-scored single-trial beta outputs. This processing yielded a dataset with 24,980 training samples and 2,770 test samples. For the test set, we averaged the three repetitions of each image, resulting in 982 test samples, but did not average the training set, following the approach of Takagi and Nishimoto [25].

We developed an individual-subject model specifically for participant 1, who completed all scanning sessions, and utilized a test set comprised of the 1,000 images that were presented to all participants. We adhered to the same standardized train/test splits used in other NSD reconstruction studies [16], [25]. Specifically, the train set for each subject contains 8,859 image stimuli and 24,980 fMRI trials. The test set includes 982 image stimuli and 2,770 fMRI trials.

3.2. Brain Decoding Pipeline

In our study, we adopted the Mind-Vis framework [6] as our baseline due to its proven efficacy in embedding fMRI data through Masked Brain Modeling (MBM). The Mind-Vis model, inspired by vision transformer architectures [8], leverages MBM to effectively capture the intricate characteristics of the fMRI data. We hypothesized that robust representation learning is essential for high-quality image reconstruction, and the Mind-Vis framework, with its innovative MBM approach, excels in this area.

The methodology involves two key stages: Sparse-Coded Masked Brain Modeling (SC-MBM) and Double-Conditioned Latent Diffusion Model (DC-LDM). Initially, MBM is pre-trained on a comprehensive dataset which includes visual areas V1 to V4 from the Human Connectome Project (HCP). This training captures the detailed features of the visual cortex, crucial for interpreting fMRI scans and reconstructing images. Vectorized voxels are divided into patches with a patch size of 16 and transformed into an embedding with 1024 dimensions. During pre-training, approximately 75% of the tokens are masked, challenging the model to reconstruct the occluded parts and enhancing its ability to generalize across diverse brain activities.

In the second stage, the trained MBM is integrated with a Latent Diffusion Model (LDM) for the conditional synthesis of images. The LDM operates on the image latent space, using the learned fMRI representations to guide the image generation process. This integration was achieved through a double-conditioning mechanism, where the fMRI latent representations were used to condition the cross-attention layers and time-step embeddings within the LDM. This dual conditioning ensured that the generated images maintained high semantic accuracy and visual fidelity.

For our model, we fine-tuned both the MBM and LDM using the Natural Scenes Dataset (NSD). This dataset includes high-resolution fMRI scans corresponding to a wide array of natural scenes, providing a rich source of data for training. Custom dataloaders were implemented to handle the NSD data, ensuring efficient loading and preprocessing. These dataloaders were designed to accommodate the specific requirements of the NSD, including handling the high-dimensional fMRI data and the corresponding image data for effective training and evaluation of the model.

3.3. Evaluation Metrics

Both quantitative and qualitative evaluations were conducted in our experiments. For quantitative evaluation, we employed nine metrics for high-level and low-level evaluation following established research. Specifically, high-level metrics included InceptionV3, CLIP, EffNet-B, SwAV,

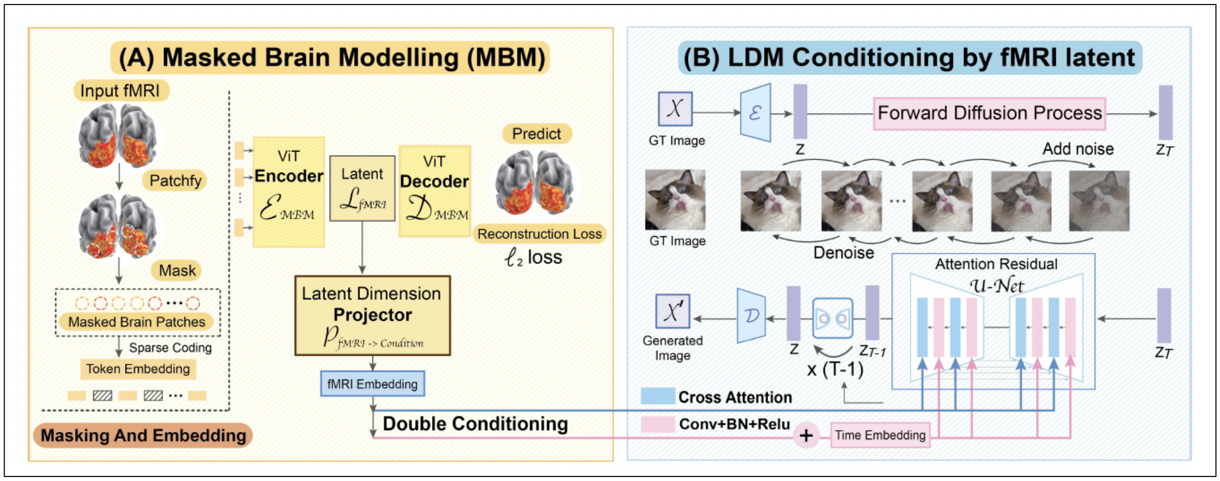

Figure 1. Model architecture of MinD-Vis

and N-way top-k accuracy, which assess various aspects of the semantic and structural fidelity of the generated images. Low-level metrics included AlexNet (2), AlexNet (5), Structural Similarity Index Measure (SSIM) and pixel-wise correlation (PixCorr), which evaluate the image quality based on pixel-level information.

PixCorr evaluates the pixel-wise correlation between the generated image and the ground truth image, providing a measure of the linear relationship between the corresponding pixels in two images. Higher correlation values indicate greater similarity.

SSIM [27] measures the structural similarity between the generated and ground truth images, considering luminance, contrast, and structure. This metric is crucial for assessing the perceptual quality of the images.

AlexNet (2) and AlexNet (5) are used to evaluate the high-level feature representations extracted from the generated images. By comparing the feature activations from different layers of AlexNet [12], we can assess how well the generated images capture the hierarchical structure of visual information.

InceptionV3 is another model used to assess the high-level semantic content of the generated images. The InceptionV3 [24] model's average pooling layer outputs are compared between the generated and ground truth images to evaluate their similarity.

CLIP (Contrastive Language-Image Pre-Training) [18] is employed to measure the alignment between the textual descriptions and the visual representations of the generated images. The model encodes both text and images into a shared latent space, enabling the comparison of their embeddings.

EffNet-B [26] is used to calculate the latent distance between the generated and ground truth images. This metric assesses the Euclidean distance between the feature vectors extracted from the generated and real images, providing insights into the high-level semantic fidelity of the generated outputs.

SwAV (Swapping Assignments between Views) [5] is a self-supervised learning method that clusters data without requiring labeled data. The SwAV metric quantifies the similarity between the neural network's internal feature representations and the brain's mechanisms for object recognition.

N-way classification accuracy [11] is used to evaluate the semantic correctness of the generated images by calculating the top-1 classification accuracy among n-1 randomly selected classes plus the correct class. This approach verifies the semantic accuracy of the generated images.

4. Experiments

4.1. SC-MBM Finetuning

We used the pre-trained models of Chen [7]. The model was trained on resting-state fMRI data from Human Connectome Project (HCP) 1200 Subject Release (600,000 fMRI segments). Hyperparameters used in the SC-MBM pre-training stage are listed in Table 1. All other unlisted parameters are set to their defaults. The SC-MBM pre-training is performed on RTX4080 SUPER GPU.

During pretraining, the model was scheduled to run for 500 epochs, with the initial 40 epochs designated as warmup epochs to gradually increase the learning rate from a lower value to the specified learning rate. The batch size was set to 100, meaning that 100 samples were processed together before the model parameters were updated. Gradient clipping was employed with a threshold of 0.8 to prevent the gradients from becoming excessively large, which can destabilize the training process. For fine-tuning, the model was trained for 30 epochs with a batch size of 32, and 2

parameter	value
patch size	16
embedding dim	1024
mask ratio	0.75
mlp ratio	1.0
decoder embed dim	512
max learning rate	2.5e-4
warm-up epochs	40
max epochs	500
encoder depth	24
encoder heads	16
decoder depth	8
decoder heads	16
clip gradient	0.8
weight decay	0.05
batch size	500
optimizer	AdamW

Table 1. Hyperparameters used in the Full model for SC-MBM Pre-training.

epochs were designated as warmup epochs. Examples of masked brain prediction are shown in Fig.3 and Fig.4.

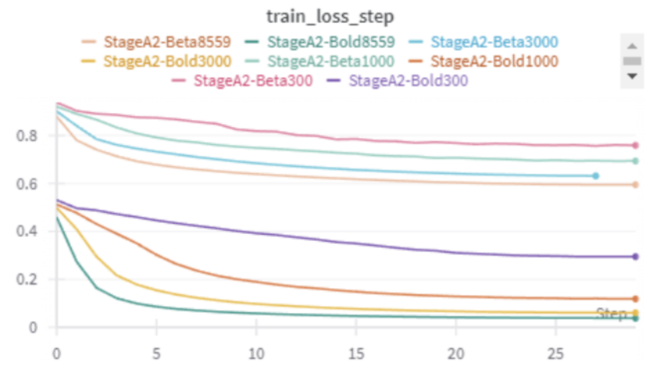

Figure 2. Loss of the model fine-tuned with SC-MBM on NSD Beta and NSD BOLD datasets. The loss is lower when using BOLD data, and more fMRI frames result in a lower loss

We observed that the loss was lower and learning occurred more rapidly with BOLD signals compared to Beta signals (Fig. 2) Considering that the number of data points in the pre-trained model was smaller than that in the target dataset, we increased the size of the target dataset and found that the loss in SC-MBM fine-tuning decreased as the amount of fMRI data increased The better performance with BOLD signals can be attributed to the difference in information content between Beta and BOLD signals, as well as the fact that the pre-trained fMRI encoder was trained on BOLD signals.

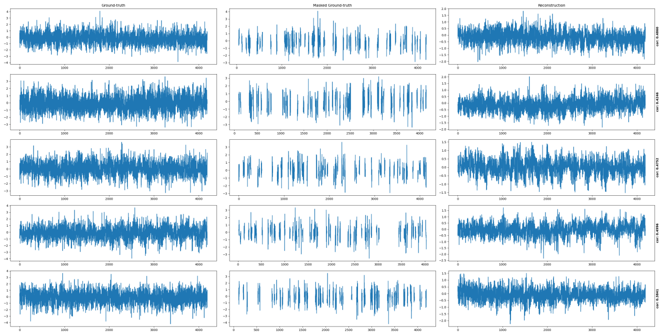

Figure 3. Examples of masked brain prediction using fMRI-beta data.First column: original fMRI data (Visual Cortex) flattened; Second column: masked fMRI; Third column: data recovered from SC-MBM decoder. Mask ratio: 0.75. The correlations between the original and recovered fMRI are also shown.

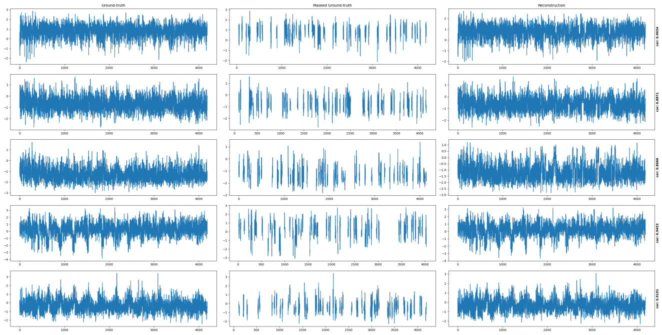

Figure 4. Examples of masked brain prediction using fMRI-BOLD data. First column: original fMRI data (Visual Cortex) flattened; Second column: masked fMRI; Third column: data recovered from SC-MBM decoder. Mask ratio: 0.75. The correlations between the original and recovered fMRI are also shown.

4.2. DC-LDM Finetuning

All fine-tunings in our experiments were performed with a single RTX 3090 Ti GPU for 50 epochs. Due to GPU memory constraints, the batch size was set to 4. We conducted fine-tuning separately on the NSD-beta and -BOLD datasets. The detailed hyperparameters are shown in Table 3.

5. Conclusion and Discussion

In this study, we investigated whether BOLD signals outperform Beta values for image reconstruction tasks with fMRI data. By comparing the performance of models using NSD's fMRI BOLD and fMRI Beta data, we aimed to demonstrate that BOLD signals capture the brain's complex and nonlinear interactions more effectively, leading to superior image reconstruction accuracy.

However, our experimental results are limited in addressing the research question due to certain methodological issues. Notably, our main model was not trained for as many epochs as reported by [6], which likely resulted in insufficient training for optimal image reconstruction. Consequently, the performance presented in this study may not be adequate to interpret the results fully.

Table 2. Comparison of Four Methods on NSD Data for Subject 1 Using Various Evaluation Metrics. The metrics used are: PixCorr (Pixel Correlation), SSIM (Structural Similarity Index), Alex(2) and Alex(5) (AlexNet top-2 and top-5 accuracy), Incep (Inception Score), CLIP (Contrastive Language-Image Pre-Training), EffNet-B (EfficientNet-B score), SwAV (Swapping Assignments between Views), and N-way Top-k classification accuracy. (For PixCorr, SSIM, AlexNet(2), AlexNet(5), Inception, N-way Top-k and CLIP metrics, higher is better. For EffNet-B and SwAV distances, lower is better. This is indicated by the arrow pointing up or down, respectively) Training epochs: Brain-Diffuser [16]: 50,000 iterations for CLIP-Text and CLIP-Vision regression, Mind-Eye [20]: 120 epoch

NSD Data	**Low-Level**				**High-Level**				
Method	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	EffNet-B ↓	SwAV ↓	N-way ↑
Takagi et al. [25]	–	–	83.0%	83.0%	76.0%	77.0%	–	–	–
Brain-Diffuser [16]	**.343**	**.372**	**92.9%**	90.6%	69.4%	64.8%	.937	.620	16.8%
Mind-Eye [20]	.205	.306	91.8%	**96.9%**	**94.6%**	**94.9%**	**.636**	**.366**	**63.6%**
Mind-Vis [6] -Beta	.063	.292	54.4%	58.9%	60.8%	57.7%	.962	.607	7.0%
Mind-Vis [6] -BOLD	.057	.288	53.3%	53.8%	53.7%	53.0%	.984	.624	4.9%

parameter	value
Epoch	50
batch size	4
image resolution	256×256×3
diffusion steps	1000
optimizer	AdamW
image latent dim	64×64×3
pre-trained type	Label-to-Image
learning rate	5.3e-5
M	77

Table 3. Hyperparameters used in the Full model for DC-LDM Finetuning.

Even if such methodological challenges are resolved, it seems likely that BOLD signals would still not perform better than Beta values in image reconstruction tasks with the currently available models. We used BOLD data that had been transformed from a 4D signal into a 2D shape. The Beta data we utilized represent the activation levels in the early and higher visual regions (i.e., nsdgeneral ROI mask). For a fair comparison, we: 1) used BOLD signals only from the visual regions of the whole brain, and 2) adjusted the 4D time series data (3D brain * timepoint) into a 2D shape. This allowed us to use the same brain ROIs as the Beta data and maintain data complexity, but likely excluded the spatio-temporal dynamics inherent in the BOLD signals.

When processing images, various brain regions within the visual network, including the visual cortex (V1 to V4) and the occipito-temporal boundary responsible for visual attention and object localization, interact in a complex manner. The fMRI BOLD signals obtained from this process include interactions between these different brain regions, unlike the Beta values derived from GLM fitting for each voxel. However, in our study, we limited the use of BOLD data to the anatomical visual cortex, which likely prevented the model from fully learning the spatial dynamics contained in the BOLD signals.

Additionally, in real-world environments, the brain processes visual stimuli in the form of videos rather than static images. The brain flexibly adapts over time, handling visual stimuli that are presented continuously. The fMRI BOLD signals obtained in this context are well-suited to reflect temporal dynamics. For instance, in the case of the NSD, images were presented for approximately 4 seconds, during which 3 fMRI frames were captured (TR = 1.33). This means that while GLM-derived Beta values yield one value per image, BOLD signals provide three-time points per image. Despite the GLM's limitations in reflecting the brain's real-time temporal dynamics, existing studies using NSD have shown good performance in image reconstruction, likely because the NSD experiments involved viewing static images without any tasks. This suggests that cognitive functions beyond simple visual processing were not engaged. In real-world settings, we do not merely process visual stimuli; we recognize objects or people, understand their purposes, and make decisions based on this information. The NSD dataset, which does not capture cognitive processes beyond basic visual processing, has demonstrated good performance in reconstructing static images. However, models trained on this dataset may struggle with more complex visual tasks.

Therefore, future fMRI-to-image reconstruction research should focus on developing models that can effectively reflect the spatio-temporal dynamics of the brain, based on data that closely mimic real-world visual processing conditions.

References

[1] Arvind Abrol, Zhihao Fu, Muhammad Salman, Ricardo Silva, Yong Du, Sergey Plis, and Vince Calhoun. Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning. *Nature communications*, 12(1):353, 2021. 2

[2] Emily J. Allen, Ghislain St-Yves, Yihan Wu, Jesse L. Breedlove, Jacob S. Prince, Logan T. Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, J. Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7t fmri dataset to bridge cognitive neuroscience and ar-

tificial intelligence. *Nature Neuroscience*, 25(1):116–126, January 2022. 1, 2, 3

[3] Roman Beliy, Guy Gaziv, Assaf Hoogi, Francesca Strappini, Tal Golan, and Michal Irani. From voxels to pixels and back: Self-supervision in natural-image reconstruction from fmri. In *Advances in Neural Information Processing Systems*, volume 32, 2019. 2

[4] Michael Breakspear. Dynamic models of large-scale brain activity. *Nature Neuroscience*, 20(3):340–352, 2017. 1

[5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020. 4

[6] Z. Chen, J. Qing, T. Xiang, W. L. Yue, and J. H. Zhou. Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22710–22720. IEEE, 2023. 1, 2, 3, 5, 6

[7] Z. Chen, J. Qing, and J. H. Zhou. Cinematic mindscapes: High-quality video reconstruction from brain activity. In *Advances in Neural Information Processing Systems*, volume 36. NeurIPS, 2024. 2, 4

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[9] Daniel J Felleman and David C Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1(1):1–47, 1991. 2

[10] Yusuke Fujiwara, Yoichi Miyawaki, and Yukiyasu Kamitani. Modular encoding and decoding models derived from bayesian canonical correlation analysis. *Neural computation*, 25(4):979–1005, 2013. 2

[11] Guy Gaziv, Roman Beliy, Niv Granot, Assaf Hoogi, Francesca Strappini, Tal Golan, and Michal Irani. Self-supervised natural image reconstruction and large-scale semantic classification from brain activity. *NeuroImage*, 254:119121, 2022. 4

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 4

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, Lecture Notes in Computer Science, pages 740–755, Cham, 2014. Springer International Publishing. 3

[14] Yoichi Miyawaki, Hajime Uchida, Okito Yamashita, Masaaki Sato, Yusuke Morito, Hiroki C Tanabe, Norihiro Sadato, and Yukiyasu Kamitani. Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, 60(5):915–929, 2008. 2

[15] Thomas Naselaris, Ryan J Prenger, Kendrick N Kay, Michael Oliver, and Jack L Gallant. Bayesian reconstruction of natural images from human brain activity. *Neuron*, 63(6):902–915, 2009. 2

[16] Furkan Ozcelik and Rufin VanRullen. Natural scene reconstruction from fmri signals using generative latent diffusion. *Scientific Reports*, 13(1):15666, 2023. 1, 3, 6

[17] Jacob S. Prince, Ian Charest, Jan W. Kurzawski, John A. Pyles, Michael J. Tarr, and Kendrick N. Kay. Improving the accuracy of single-trial fmri response estimates using glmsingle. *eLife*, 11:e77599, November 2022. 3

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 4

[19] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[20] Paul Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Aidan Dempster, Nathalie Verlinde, et al. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 3, 6

[21] Guohua Shen, Kshitij Dwivedi, Kei Majima, Tomoyasu Horikawa, and Yukiyasu Kamitani. End-to-end deep image reconstruction from human brain activity. *Frontiers in Computational Neuroscience*, 13, 2019. 2

[22] Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. Deep image reconstruction from human brain activity. *PLoS computational biology*, 15(1):e1006633, 2019. 2

[23] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[24] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016. 4

[25] Y. Takagi and S. Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14453–14463, 2023. 1, 2, 3, 6

[26] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019. 4

[27] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 4