Multimodal decoding of human brain activity into images and text

Matteo Ferrante
Department of Biomedicine and Prevention
University of Rome, Tor Vergata
matteo.ferrante@uniroma2.it

Tommaso Boccato
Department of Biomedicine and Prevention
University of Rome, Tor Vergata

Furkan Ozcelik
CerCo, CNRS UMR5549, Toulouse, France
Universite de Toulouse, Toulouse, France
ANITI, Toulouse, France

Rufin VanRullen
CerCo, CNRS UMR5549, Toulouse, France
Universite de Toulouse, Toulouse, France
ANITI, Toulouse, France

Nicola Toschi
Department of Biomedicine and Prevention
University of Rome, Tor Vergata
Martinos Center For Biomedical Imaging
MGH and Harvard Medical School (USA)

Editors: Marco Fumero, Emanuele Rodolà, Clementine Domine, Francesco Locatello, Gintare Karolina Dziugaite, Mathilde Caron

Abstract

Every day, the human brain processes an immense volume of visual information, relying on intricate neural mechanisms to perceive and interpret these stimuli. Recent breakthroughs in functional magnetic resonance imaging (fMRI) have enabled scientists to extract visual information from human brain activity patterns. In this study, we present an innovative method for decoding brain activity into meaningful images and captions, with a specific focus on brain captioning due to its enhanced flexibility as compared to brain decoding into images. Our approach takes advantage of cutting-edge image captioning models and incorporates a unique image reconstruction pipeline that utilizes latent diffusion models and depth estimation.
We utilized the Natural Scenes Dataset, a comprehensive fMRI dataset from eight subjects who viewed images from the COCO dataset. We employed the Generative Image-to-text Transformer (GIT) as our backbone for captioning and propose a new image reconstruction pipeline based on latent diffusion models. The method involves training regularized linear regression models between brain activity and extracted features. Additionally, we incorporated depth maps from the ControlNet model to further guide the reconstruction process.
We propose a multimodal based approach that leverages similarities between neural and deep learning representations and by learning alignment between these spaces, we produce textual description and image reconstruction from brain activity.
We evaluate our methods using quantitative metrics for both generated captions and images. Our brain captioning approach outperforms existing methods, while our image reconstruction pipeline generates plausible images with improved spatial relationships.
In conclusion, we demonstrate significant progress in brain decoding, showcasing

Proceedings of the I edition of the Workshop on Unifying Representations in Neural Models (UniReps 2023).

the enormous potential of integrating vision and language to better understand human cognition. Our approach provides a flexible platform for future research, with potential applications based on a combination of high-level semantic information coming from text and low-level image shape information coming from depth maps and initial guess images.

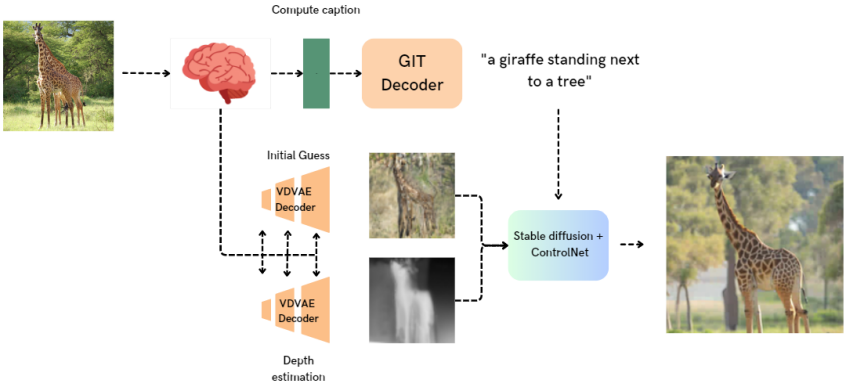

Figure 1: Our model utilizes fMRI measurements to extract features for GIT captioning and VDVAE initial and depth image estimation using linear models. Image captions serve as the primary general result, used in the second stage alongside other conditioning to generate plausible reconstructions with a latent diffusion model. GIT and VDVAE models are pre-trained and frozen, while linear regressions are trained from fMRI to their latent spaces.

1 Introduction

The human visual system is an extraordinary product of evolution, enabling us to navigate and interact with our surroundings. From basic patterns to intricate scenes, our brains persistently process and interpret visual information. A central challenge in neuroscience is comprehending how these elaborate processes occur at the neural activity level. Functional magnetic resonance imaging (fMRI) has emerged as an essential tool for studying neural activity associated with visual perception, by measuring blood oxygen level-dependent (BOLD) signals. Brain decoding has progressed significantly, employing fMRI data to reconstruct visual stimuli from brain activity patterns. This has the potential to revolutionize our understanding of the neural code underlying visual perception with possible applications in brain-computer interfaces and clinical diagnostics. The increasing interest in reconstructing information from noninvasive brain data is driven by enhanced data availability, improved computational power, and sophisticated deep learning methods. Despite challenges with signal-to-noise ratio, session duration, and hemodynamic response function variability, fMRI has proven effective in various tasks such as visual stimulus and text classification and reconstruction [Schneider et al., Zafar et al., 2015, Lindsay, 2021, Awangga et al., 2020].

In this work, our first contribution is shifting the prediction from images to text, aiming to generate a caption of the observed scene from brain activity. To compare with prior work, we propose a new model for image captioning from brain activity and propose a new image reconstruction pipeline based on a conditioned and controlled version of the latent diffusion model, Stable Diffusion. Predicting a caption instead of the image in brain decoding from fMRI of visual stimuli offers several advantages. Captions naturally represent a higher level of abstraction, requiring a more advanced interpretation and summarization of visual information than merely predicting the image itself. As a result, predicting captions can help us understand how the brain processes and represents complex visual information. In real-world situations, humans often describe visual scenes with words, so predicting captions instead of images may better capture an important aspect of visual information processing. Recent neuroscience research has shown substantial evidence that large language models can be correlated with brain activity and that it is possible to predict one representation from the other [Caucheteux and King, Tang et al.]. Finally, predicting text from fMRI could lead to better generalization across modalities. Natural language is our main tool as humans to interact with each other and nowadays even with foundation models. We can exploit large language models to condition

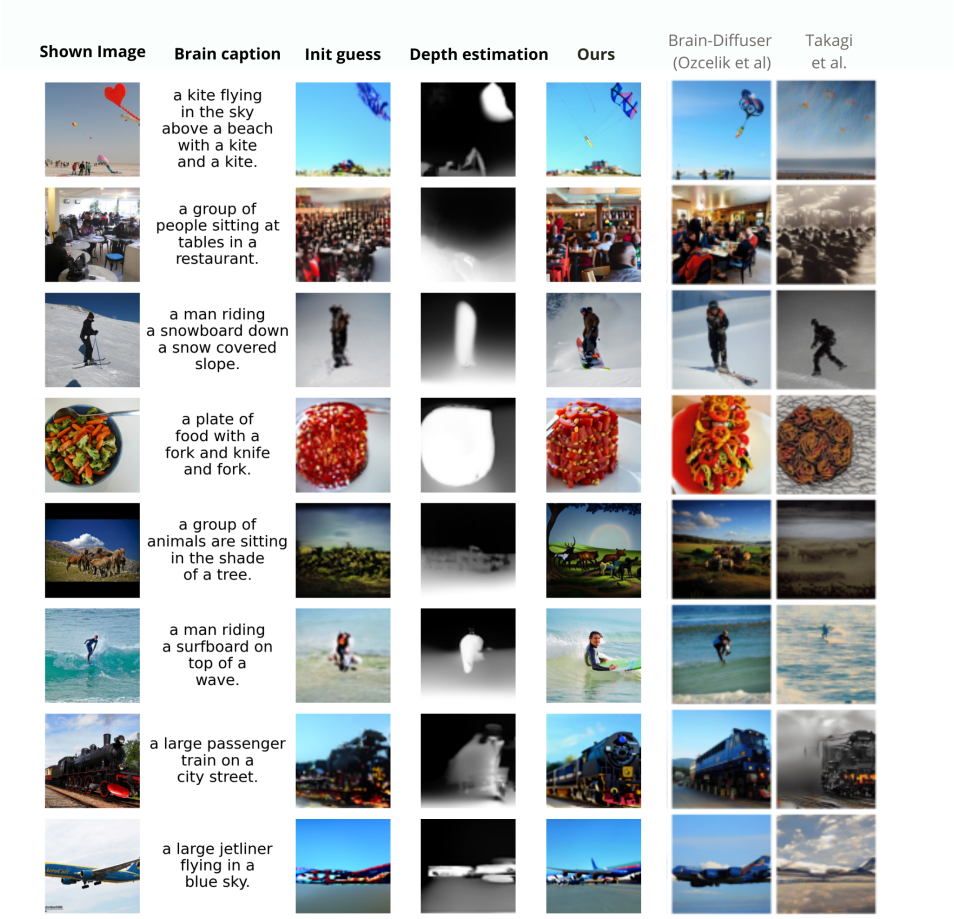

Figure 2: Comparison of our results (Columns 2-4) with the shown stimuli and reconstructions from other works. The second column displays the caption computed from the brain activity, the third column presents the initial guess image, the fourth column shows the depth estimated images, and the fifth column reports our final reconstruction. The last two columns showcase reconstructions from two recent works. All results are from subj01.

other models to generate images, videos, audio, and more. Predicting text from brain helps us rapidly change the reconstruction model, leveraging state-of-the-art text-to-image models to generate realistic images from brain activity. In summary, our contributions in this paper are two-fold: We propose a method to generate image captions from brain activity using a multimodal large language model [Wang et al.] and introduce a novel image reconstruction pipeline based on predicted text and estimated initial and depth maps from brain activity. The main novelty proposed in our work is a pipeline that leverage aligned representations of brain, text and images for visual stimuli. Fig 1 is a scheme of the entire procedure that we propose, while Fig 2 shows generated captions and images from brain activity compared to other image reconstruction methods.

1.1 Related Works

In the field of brain decoding, researchers have utilized various modeling frameworks with pre-processed fMRI time series as input. These data have served as the basis for numerous decoding approaches. Some examples include employing a variational autoencoder with a generative adversarial component (VAE-GAN) to encode latent representations of human faces [VanRullen and Reddy, 2019] and applying sparse linear regression on preprocessed fMRI data to predict features extracted from early convolutional layers in a pre-trained CNN [Horikawa and Kamitani, 2017] for natural images. Unsupervised and adversarial strategies have been used to reconstruct images, incorporating dual VAE-GAN and unsupervised methods for fMRI stimuli decoding with various encoders and decoders trained in different ways [Shen et al., 2019, Ren et al., 2019, Gaziv et al.,

Shown Image	COCO Caption	BrainCaptioner subj01	BrainCaptioner subj02
	a giraffe standing in a fenced in area.	a giraffe standing in a park.	a giraffe standing in a park.
	a man holding a surfboard in the ocean.	a person is riding a surfboard on a wave.	a man standing on a surfboard on a beach.
	a group of people riding skis down a snow covered slope.	a group of people riding skis down a snow covered slope.	a group of people riding on top of a snow covered slope.

Shown Image	COCO Caption	BrainCaptioner subj01	BrainCaptioner subj02
	a man holding a pastry in his hand.	a young man holding a bowl of food.	a young man holding a slice of pizza.
	airliner on the runway	a large passenger jet airplane on a runway.	a large white and blue plane on a runway.
	a group of people sitting at tables in a cafeteria.	a group of people sitting at tables in a restaurant.	a group of people sitting around a table.

Figure 4: Examples of generated caption with our BrainCaptioner pipeline. Shown images are test set stimuli used for subj01 and subj02 during the fmri experiment. COCO Caption column report the first annotations for the original COCO image, while the other two columns are the output of our model for the two subjects.

2022]. Optimizing the latent spaces of pretrained architectures, such as BigBiGAN and IC-GAN, can facilitate reconstructing high-quality images from fMRI patterns [Donahue and Simonyan, 2019, Casanova et al., 2021, Mozafari et al., 2020, Ozcelik et al., 2022]. Recently, diffusion models have become a significant component of the decoding pipeline due to their improved performance in image generation [Takagi and Nishimoto, 2023, Chen et al., 2022], also incorporating semantic-based strategies like [Ferrante et al., 2023] or multi-step decoding strategies as in [Ozcelik and VanRullen, 2023, Chen et al., 2023, Scotti et al., 2023, Tang et al.]. To the best of our knowledge, only a few works [Takada et al., Matsuo et al., 2016, Qiao et al., 2018] have attempted brain captioning, utilizing a combination of a pre-trained convolutional neural network and recurrent neural network for captioning and estimating the convolutional features from brain activity. The primary differences between our work and previous research are the shift in paradigm from direct image estimation to brain captioning and leveraging multimodal transformer-based language models, which have been shown to better describe brain activity [Choksi et al.].

2 Methods

In this section, we describe the proposed method and the data we used. The data are publicly available and can be requested at `https://naturalscenesdataset.org/`. All experiments and models were trained on a server equipped with four A100 GPU cards and 2 TB of RAM. The entire analysis took approximately 16 hours per subject. The pipelines are based on pre-trained versions of deep learning models used as proxies for brain activity, generating latent representations that could be similar (and thus linearly mapped) to brain activity and vice versa.

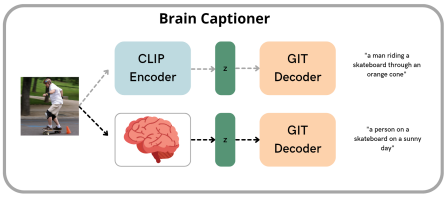

Figure 3: Image captioning from brain activity pipeline: Gray dotted lines are only used during training, and only orange boxes are used during inference, replacing their inputs with those estimated from brain activity.

Data:We employed the Natural Scenes Dataset (NSD) [Allen et al., 2022], a comprehensive fMRI dataset featuring eight subjects who viewed images from the COCO dataset. Our analysis concentrated on four subjects (same used in other decoding works for comparison), yielding a training set of 8,859 images and 24,980 fMRI trials, and a test set of 982 images and 2,770 fMRI trials per subject. Images are repeated up to three times and their trials were averaged to increase signal-to-noise ratio. To reduce spatial dimensionality to approximately 15,000 voxels, the fMRI signal (1.8mm resolution) was masked using the NSDGeneral ROI mask, which covers numerous visual areas. This ROI selection is vital for enhancing the signal-to-noise ratio and minimizing data complexity. The chosen ROI mask facilitated the investigation of both low-level and high-level visual features. To decrease temporal dimensionality, we employed precomputed betas from a GLM with fitted HRF and denoised as described in the NSD paper.

Captioning model and renormalization: For brain captioning, we utilized the state-of-the-art image captioning model, GIT [Wang et al.], as our backbone. GIT (Generative Image-to-text Transformer) is an innovative model designed to integrate vision and language tasks. In contrast to conventional

approaches that depend on intricate architectures and external modules, GIT adopts a streamlined structure consisting of a single image encoder and a text decoder, unified under one language modeling task. Leveraging large-scale pre-training data and model size, GIT outperforms existing models on 12 benchmarks and even surpasses human performance on TextCaps. Essentially, GIT comprises a CLIP Vision encoder [Radford et al., 2021] followed by a GPT decoder, trained on large-scale datasets. For the stimuli in the train set, we computed features from images and trained a regularized linear regression to map between brain activity and these features. We used cross-validation to select the best regularization parameter α and discovered that a value of 50,000 performed optimally using the negative mean squared error as a scoring function. This is our brain-to-features model, which serves as the core component of our method for brain captioning. Before feeding estimated features to the decoder, we required a normalization pass. Thus, we computed the mean and standard deviation of features from images and those predicted by the model over the training set, replacing their values during inference on the test set to match the real feature distributions. A schematic representation of the overall pipeline can be seen in Fig. 3 and generated captions from this pipeline for both subjects are shown in Fig 4.

Reconstruction pipeline: Recent research in brain decoding has focused on developing image reconstruction techniques [Ozcelik and VanRullen, 2023, Ozcelik et al., 2022, Takagi and Nishimoto, 2023, Lin et al., 2022, Chen et al., 2022]. Studies have demonstrated that high SNR fMRI data of visual stimuli enables effective brain decoding using diffusion models. Various approaches have been proposed to enhance these models' performance, with the optimal method for image reconstruction remaining an open question. One approach to improve low-level detail generation and increase the similarity between original and decoded images is to provide the network with an initial guess image or an estimated latent space.

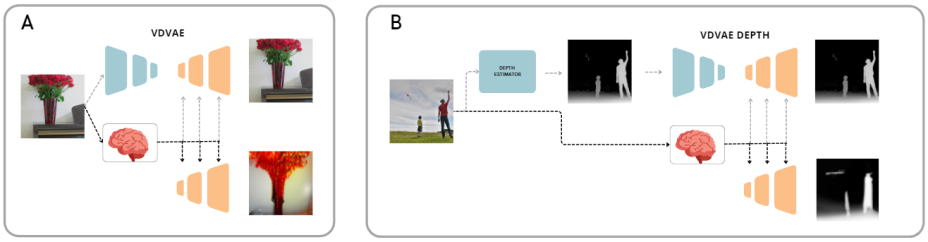

Figure 5: **A**: Pipeline for initial images capturing 2D RGB pixel information. **B**: Pipeline for inferred depth estimates. Both depth image and the initial image are estimated from brain activity. Gray dotted lines are only used during training, while only orange boxes are used during inference, replacing their inputs with the ones estimated from brain activity.

Initial Guess: To compare our approach with existing research on brain decoding, we augmented our method by proposing an image reconstruction pipeline based on latent diffusion models. Following the approach described in [Ozcelik and VanRullen, 2023], we initially estimate a "guess image" to generate an approximate initial image with colors and shapes. To achieve this, we computed the latent representations of the first 31 layers of the very deep variational autoencoder model [Child, 2021] (VDVAE), pre-trained on natural images, and kept frozen. In a VDVAE, the encoder network maps the input data onto a lower-dimensional latent space, while the decoder network maps the latent space back to the original data space. The architecture of the VAE is hierarchical. In other words, the hidden units in each layer depend not only on the input data but also on the outputs of the previous layer. This conditional dependence allows the VAE to capture complex relationships between the input data and the latent space, resulting in a more powerful and expressive model. Consequently, we trained a regularized linear regression between brain activity and estimated features for each of the first 31 layers, using the renormalization procedure described in the previous section to match the target distribution. During inference over the test set, these features are estimated from brain activity, renormalized, and passed to the VDVAE decoder to reconstruct an initial image, as depicted in Fig. 5.

Depth estimation: We propose using ControlNet [Zhang and Agrawala, 2023] to augment Stable Diffusion [Rombach et al., 2021], a state-of-the-art latent diffusion model, for improving foreground-background matching in reconstructed images by incorporating depth information. We first compute grayscale depth images for all training stimuli using Dense Vision Transformer and the Huggingface library [Ranftl et al., 2021, Wolf et al., 2019]. We then pass these depth images into the Variational

Metric	Baselines				**Ours**			
	subj01	subj02	subj05	subj07	subj01	subj02	subj05	subj07
Meteor (image vs human)	0.176	0.174	0.177	0.175	0.404	0.404	0.404	0.404
Meteor (brain vs image)	0.163	0.166	0.166	0.166	**0.305**	**0.298**	**0.303**	**0.291**
Sentence (image vs human)	0.319	0.315	0.321	0.315	0.703	0.703	0.703	0.703
Sentence (brainvs image)	0.280	0.281	0.282	0.281	**0.447**	**0.418**	**0.443**	**0.413**
CLIP (image vs human)	0.672	0.673	0.676	0.673	0.831	0.831	0.831	0.831
CLIP (brain vs image)	0.624	0.627	0.626	0.627	**0.705**	**0.688**	**0.702**	**0.693**

Table 1: Text Metrics Comparison: This table reports the values of various metrics for each subject, both for the baseline and our model (columns). Each row represents a different metric. Metrics labeled with "(image captions and human captions)" evaluate the model-generated captions from images against the original COCO captions, serving as a comparison of the model's performance. Metrics labeled with "(brain captions and image captions)" pertain to captions computed from brain activity.

Diffusion Autoencoder (VDVAE) model and train a regularized linear regression from brain activity to the model's latent, as illustrated in Fig 5.The VDVAE is the same used before (pre-trained on natural images and kept frozen), however here it is here to generate latent representation of the estimated depth images, which are our target for regression.

Whole Reconstruction pipeline: The pipeline (Fig 1) first decodes brain activity into a latent space to generate captions for test stimuli using learned ridge regression. Then, the initial guess and depth images are computed from brain activity to condition the latent model. Stable Diffusion v2 + ControlNet is used for implementation, with 30 inference steps, guidance scale 9, and control net weight 0.8. The negative prompt sentence *i*s also included to improve quality.

Evaluation: We compared our brain captioning work with existing methods by re-implementing the architecture from [Takada et al.], consisting of a CNN followed by an LSTM. We used Ridge regression to map brain activity to the CNN's final convolutional layer and applied renormalization before feeding the LSTM. We evaluated the generated captions using metrics such as METEOR, CLIP similarity, and SentenceTransformer similarity. Additionally, we assessed our image reconstruction pipeline using low-level and high-level metrics like PixCorr, SSIM, 2-way accuracy in AlexNet, Inception, and CLIP latent spaces, and FID, allowing comparison with other brain decoding studies.

3 Results

Table 3 presents the results of the evaluation of the proposed approach compared to the baseline models and previous works. This table reports text-based metrics, including Meteor score, CLIP, and SentenceTransformer similarity, computed for the reference captions, captions generated from images by both models (baseline and proposed), and captions generated from brain activity using the proposed approach. Results show that our approach outperforms the baseline models on all metrics and achieves significantly higher scores than previous works, indicating the effectiveness of the approach in generating accurate and meaningful captions from brain activity.
The table 2 reports image-based metrics, including PixCorr, SSIM, accuracy in various layers of AlexNet and Inception, CLIP similarity, and FID score. Results show that the proposed approach outperforms the previous works in low-level metrics, including PixCorr, SSIM and the lower layer of AlexNet. High level metrics are on par or slightly lower than state-of-the-art methods, probably due to a bottleneck in text predictions. If a word is predicted wrongly, this error is propagated in the image reconstruction pipeline and impacts on high-level metrics. Overall, the results demonstrate the effectiveness of the proposed approach in decoding brain activity into meaningful images and captions, performing on par on even outperforming state-of-the-art in several metrics. Fig 2, 6, 4 and figures in the supplementary material show some visual comparison with other works for a qualitative comparison. Qualitatively, the captions represent plausible descriptions of images matching the high-level semantic content in most of cases. Sometimes, captions are more general with descriptions like "animals in the grass" instead of the specific type of animal. In other cases, only details are missing (or wrong). For example, in Fig 4 for the surfer image for one subject, the model adds "on a wave" while for the other the model specifies "on a beach". Similarly, in the first image of the right part, the pastry in the man's hand is changed to "a bowl of foods" or "slice of pizza". This could

Model	Low level metrics			High level		
	PixCorr	SSIM	AlexNet (2)	AlexNet (5)	Inception	CLIP
Lin et al (2022)	-	-	-	-	0.782	-
Takagi et al (2022)	-	-	0.83	0.83	0.76	0.77
Gu et al (2023)	0.15	0.325	-	-	-	-
Ozcelik et al (2023)	0.30	0.28	**0.89**	**0.98**	**0.92**	**0.94**
Our Model	**0.353**	**0.327**	**0.89**	0.97	0.84	0.90

Table 2: Image Metrics Analysis: Metrics from Ozcelik et al were recomputed by requesting images from subj01 and subj02 from the authors and averaging them to facilitate comparison with our results. Metrics from other works are cited directly from the original articles.

support the hypothesis that our pipeline is able to capture the main characteristic of the images from brain activity and the GIT decoder help in plausible sentence decoding.

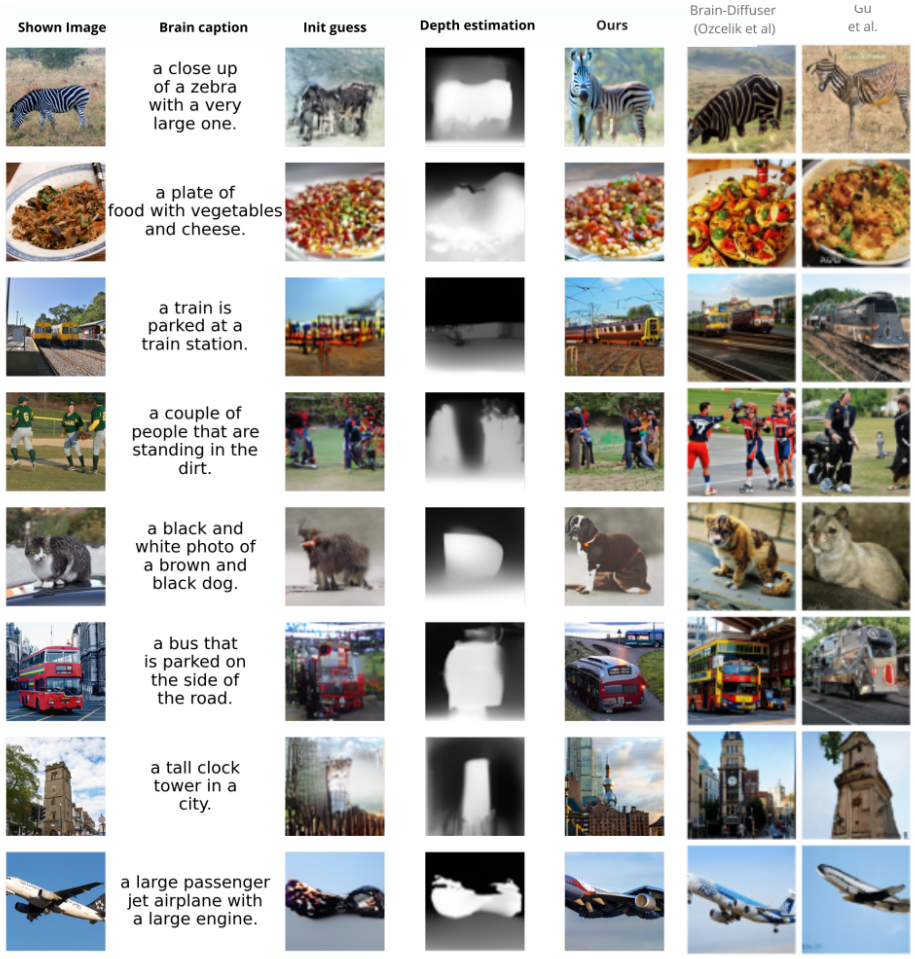

Figure 6: Comparison of our results (Columns 2-4) with the presented stimuli and other reconstruction works. The second column displays the caption derived from brain activity, the third column presents the initial guess image, the fourth column exhibits the depth-estimated images, and the fifth column showcases our final reconstruction. The last two columns demonstrate reconstructions from two recent works. All results are from subj01.

4 Discussion

In this study, we proposed a method to generate captions from brain activity measured during a vision task. The primary motivation for shifting from image reconstruction to image captions is the

flexibility of manipulating text prompts and the ease of modifying the image reconstruction pipeline as separate modules. We also proposed an image reconstruction pipeline that incorporates depth maps and initial guesses to generate plausible images. Depth maps provide information about the spatial relationships between objects in a scene injecting information that could improve the overall quality of the reconstructed images.

Neural Art and Examples: Our approach has potential applications in neural art and style transfer. By leveraging our image reconstruction pipeline, we can explore the creative space of combining content and style from different text prompts. This could lead to the generation of visually captivating art, expanding the possibilities for artistic expression using AI. For example, modifying inputs by adding specific styles could drive the diffusion process toward an image with the same content but a different style. This approach represents a novel type of art that combines artificial intelligence, neuroscience, and creativity, starting from the decoded activity of the brain that could be modulated by a text description of the scene.

Ethics: As brain decoding research advances, ethical considerations must be addressed. For instance, the potential misuse of image reconstruction and generative models to create misleading or harmful content raises concerns, given that decoded activity is related to the mental and internal states of someone. It is crucial to develop guidelines and policies that ensure responsible use and prevent the exploitation of this technology for malicious purposes. Additionally, we must consider potential biases in the training data, as these can propagate and influence the generated output, perpetuating stereotypes and unfair representations, unrelated to thoughts of the specific subject. There are also possible concerns about privacy, given that brain decoding models are able to decode language, thoughts, and perceptions [Schneider et al., Tang et al.]. From early experiments, it seems that high-level performances are only achievable when subjects are collaborating because the attention process can warp [Çukur et al., 2013] the semantic representation in the brain, which is the primary target of these deep learning multimodal models used as a proxy for brain activity [Choksi et al.].
Limitations: In our investigation of brain decoding, we have identified several key limitations that impact the efficacy and generalizability of our findings. The following discussion aims to elaborate on these constraints, establish their interconnections, and provide a deeper understanding of the challenges we face in advancing this field of research. A major limitation in brain decoding work is the necessity for subject-specific models. Individual differences in brain structure, function, and cognitive processing make it challenging to develop a universal decoding model. This specificity hinders the broader applicability of our findings and demands the development of personalized models for each subject.
Even for subject-specific models, to achieve reliable and accurate decoding, a significant amount of high-quality data is needed. Obtaining such data is often time-consuming and resource-intensive, limiting the scalability of brain decoding studies. Additionally, low SNR data can introduce errors and inconsistencies in the decoding process, further compromising the reliability of the results. In this work we used a 7T dataset, that inherently has higher SNR with respect to previous 3T datasets [Horikawa and Kamitani, 2017], enhancing the quality of our results. In our work, the image captioning model acts as an upper limit: the performance of our brain captioning pipeline is inherently limited by the GIT image captioning model employed. Any inaccuracies or biases present in the model will directly impact the quality of decoded information, setting an upper bound on the performance that can be achieved. Also, the quality of the mapping between neural activity and external stimuli representation in latent spaces is another critical factor influencing the performance of our approach. This determines the accuracy and resolution of the decoded information. Current methods, however, are often limited by the complexity and variability of brain activity, as well as the constraints imposed by the data acquisition techniques, and usually rely on simple regression techniques. Addressing these challenges is essential for refining the mapping process and improving decoding outcomes. Regarding image reconstruction, generating images from text could be another bottleneck. If the text contains errors, these will be propagated and/or enhanced by a separate image reconstruction pipeline. This represents the price for increased flexibility and independence from the specific image reconstruction pipeline used. Finally, the brain decoding process may involve multiple areas, including temporal poles, which further impact of performances. Different brain regions may process and represent information differently, and understanding these variations is crucial for developing accurate and comprehensive decoding models. With the aim of reducing spatial dimensionality, we used only a visual responding region defined by the NSDGeneral ROI, however other brain areas could also encode relevant pieces of information that are relevant to

improve performances. Exploring performances as a function of different input regions could be an interesting field of future research.

5 Conclusions

Our approach builds upon neuroscientific and AI concepts, leveraging multimodal models to generate captions from brain activity related to the vision of different scenes. We augmented our brain captioning with a pipeline for image reconstruction that uses predicted text and initial information about colors and depth also estimated by brain activity. In conclusion, our approach demonstrates promising results in image captioning and reconstruction from brain activity, with potential applications in a number of cross-disciplinary fields. By drawing on these foundations, we could further our understanding of the human brain's processing of visual and language information, ultimately improving related AI algorithms as well as applications. As we refine our approach, we can continue to explore the intricate relationship between neuroscience and AI, potentially uncovering novel insights and fostering interdisciplinary collaboration.

References

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 artificial intelligence. *Nature Neuroscience*, 25(1):116–126, Jan 2022. ISSN 1546-1726. doi: 10.1038/s41593-021-00962-x. URL https://doi.org/10.1038/s41593-021-00962-x.

R M Awangga, T L R Mengko, and N P Utama. A literature review of brain decoding research. *IOP Conference Series: Materials Science and Engineering*, 830(3):032049, April 2020. ISSN 1757-8981, 1757-899X. doi: 10.1088/1757-899X/830/3/032049. URL https://iopscience.iop.org/article/10.1088/1757-899X/830/3/032049.

Arantxa Casanova, Marlène Careil, Jakob Verbeek, Michal Drozdzal, and Adriana Romero-Soriano. Instance-conditioned gan, 2021. URL https://arxiv.org/abs/2109.05070.

Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. 5(1):134. ISSN 2399-3642. doi: 10.1038/s42003-022-03036-1. URL https://www.nature.com/articles/s42003-022-03036-1.

Zijiao Chen, Jiaxin Qing, Tiange Xiang, Wan Lin Yue, and Juan Helen Zhou. Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding, 2022.

Zijiao Chen, Jiaxin Qing, Tiange Xiang, Wan Lin Yue, and Juan Helen Zhou. Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding, 2023.

Rewon Child. Very deep vaes generalize autoregressive models and can outperform them on images, 2021.

Bhavin Choksi, Milad Mozafari, Rufin VanRullen, and Leila Reddy. Multimodal neural networks better explain multivoxel patterns in the hippocampus. 154:538–542. ISSN 0893-6080. doi: 10.1016/j.neunet.2022.07.033. URL https://www.sciencedirect.com/science/article/pii/S0893608022002982.

Jeff Donahue and Karen Simonyan. Large scale adversarial representation learning, 2019. URL https://arxiv.org/abs/1907.02544.

Matteo Ferrante, Tommaso Boccato, and Nicola Toschi. Semantic brain decoding: from fmri to conceptually similar image reconstruction of visual stimuli, 2023.

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, July 2022. ISSN 10538119. doi: 10.1016/j.neuroimage.2022.119121. URL https://linkinghub.elsevier.com/retrieve/pii/S105381192200249X.

Katherine L. Hermann, Ting Chen, and Simon Kornblith. The origins and prevalence of texture bias in convolutional neural networks, 2020.

Tomoyasu Horikawa and Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, 8(1):15037, August 2017. ISSN 2041-1723. doi: 10.1038/ncomms15037. URL http://www.nature.com/articles/ncomms15037.

Sikun Lin, Thomas Sprague, and Ambuj K Singh. Mind reader: Reconstructing complex images from brain activities, 2022.

Grace W Lindsay. Convolutional neural networks as a model of the visual system: Past, present, and future. *J. Cogn. Neurosci.*, 33(10):2017–2031, September 2021.

Eri Matsuo, Ichiro Kobayashi, Shinji Nishimoto, Satoshi Nishida, and Hideki Asoh. Generating Natural Language Descriptions for Semantic Representations of Human Brain Activity. In *Proceedings of the ACL 2016 Student Research Workshop*, pages 22–29, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-3004. URL http://aclweb.org/anthology/P16-3004.

Milad Mozafari, Leila Reddy, and Rufin VanRullen. Reconstructing Natural Scenes from fMRI Patterns using BigBiGAN. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, July 2020. doi: 10.1109/IJCNN48605.2020.9206960. URL http://arxiv.org/abs/2001.11761. arXiv:2001.11761 [cs, eess, q-bio].

Furkan Ozcelik and Rufin VanRullen. Brain-diffuser: Natural scene reconstruction from fmri signals using generative latent diffusion, 2023.

Furkan Ozcelik, Bhavin Choksi, Milad Mozafari, Leila Reddy, and Rufin VanRullen. Reconstruction of Perceived Images from fMRI Patterns and Semantic Brain Exploration using Instance-Conditioned GANs, February 2022. URL http://arxiv.org/abs/2202.12692. arXiv:2202.12692 [cs, eess, q-bio].

Kai Qiao, Chi Zhang, Linyuan Wang, Jian Chen, Lei Zeng, Li Tong, and Bin Yan. Accurate Reconstruction of Image Stimuli From Human Functional Magnetic Resonance Imaging Based on the Decoding Model With Capsule Network Architecture. *Frontiers in Neuroinformatics*, 12: 62, September 2018. ISSN 1662-5196. doi: 10.3389/fninf.2018.00062. URL https://www.frontiersin.org/article/10.3389/fninf.2018.00062/full.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction, 2021.

Ziqi Ren, Jie Li, Xuetong Xue, Xin Li, Fan Yang, Zhicheng Jiao, and Xinbo Gao. Reconstructing Perceived Images from Brain Activity by Visually-guided Cognitive Representation and Adversarial Learning, October 2019. URL http://arxiv.org/abs/1906.12181. arXiv:1906.12181 [cs].

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. URL https://arxiv.org/abs/2112.10752.

Steffen Schneider, Jin Hwa Lee, and Mackenzie Weygandt Mathis. Learnable latent embeddings for joint behavioural and neural analysis. pages 1–9. ISSN 1476-4687. doi: 10.1038/s41586-023-06031-6. URL https://www.nature.com/articles/s41586-023-06031-6. Publisher: Nature Publishing Group.

Paul S. Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Ethan Cohen, Aidan J. Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth A. Norman, and Tanishq Mathew Abraham. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors, 2023.

Guohua Shen, Kshitij Dwivedi, Kei Majima, Tomoyasu Horikawa, and Yukiyasu Kamitani. End-to-end deep image reconstruction from human brain activity. *Front. Comput. Neurosci.*, 13:21, April 2019.

Saya Takada, Ren Togo, Takahiro Ogawa, and Miki Haseyama. Generation of viewed image captions from human brain activity via unsupervised text latent space. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 2521–2525. doi: 10.1109/ICIP40778.2020.9191262. ISSN: 2381-8549.

Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. *bioRxiv*, 2023. doi: 10.1101/2022.11.18.517004. URL https://www.biorxiv.org/content/early/2023/03/11/2022.11.18.517004.

Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G. Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. 26(5):858–866. ISSN 1546-1726. doi: 10.1038/s41593-023-01304-9. URL https://www.nature.com/articles/s41593-023-01304-9. Number: 5 Publisher: Nature Publishing Group.

Rufin VanRullen and Leila Reddy. Reconstructing faces from fmri patterns using deep generative neural networks. *Communications Biology*, 2(1):193, May 2019. ISSN 2399-3642. doi: 10.1038/s42003-019-0438-y. URL https://doi.org/10.1038/s42003-019-0438-y.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. GIT: A generative image-to-text transformer for vision and language. URL http://arxiv.org/abs/2205.14100.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2019. URL https://arxiv.org/abs/1910.03771.

Raheel Zafar, Aamir Saeed Malik, Nidal Kamel, Sarat C. Dass, Jafri M. Abdullah, Faruque Reza, and Ahmad Helmy Abdul Karim. Decoding of visual information from human brain activity: A review of fMRI and EEG studies. *Journal of Integrative Neuroscience*, 14(02):155–168, June 2015. ISSN 0219-6352, 1757-448X. doi: 10.1142/S0219635215500089. URL http://www.worldscientific.com/doi/abs/10.1142/S0219635215500089.

Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

Tolga Çukur, Shinji Nishimoto, Alexander G Huth, and Jack L Gallant. Attention during natural vision warps semantic representation across the human brain. *Nature Neuroscience*, 16(6):763–770, June 2013. ISSN 1097-6256, 1546-1726. doi: 10.1038/nn.3381. URL http://www.nature.com/articles/nn.3381.

6 Supplementary Material

In this section, more comparisons of captions and reconstructed images are provided, compared with state-of-the-art brain decoding pipelines.

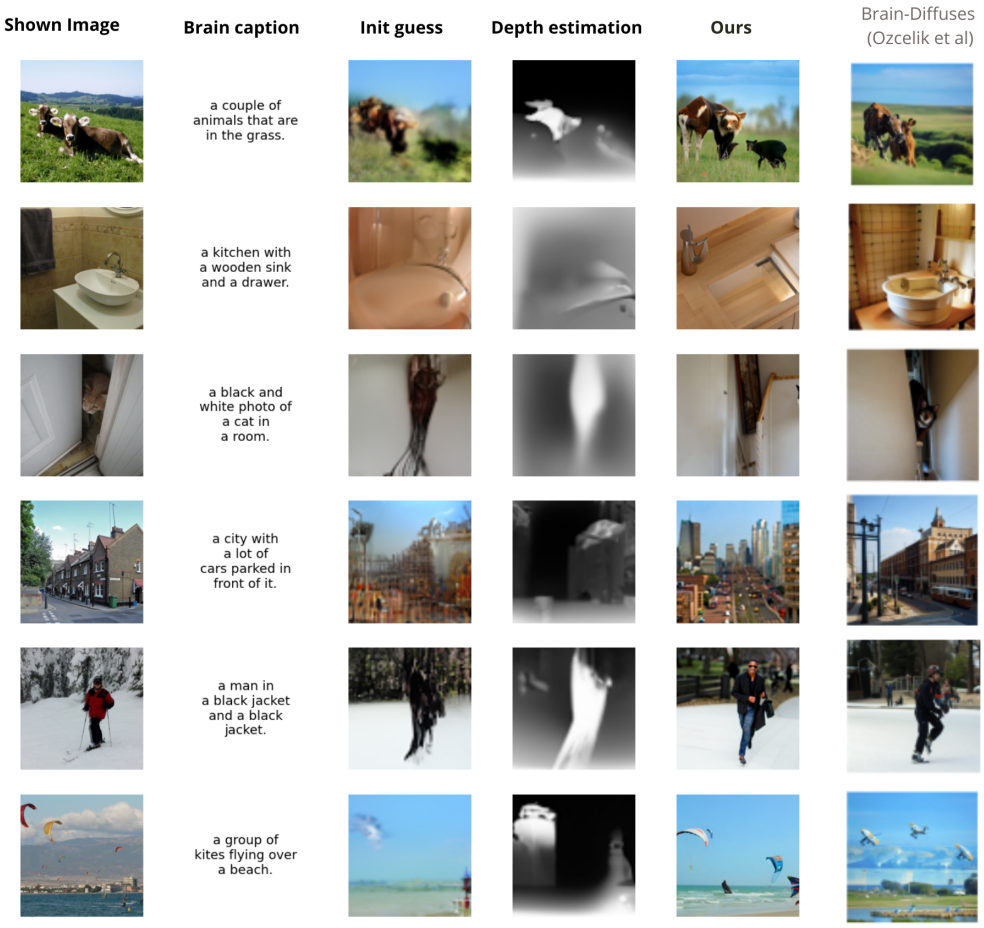

Figure A1: Comparison of our results (Columns 2-4) with the presented stimuli and other reconstruction works. The second column displays the caption derived from brain activity, the third column presents the initial guess image, the fourth column exhibits the depth-estimated images, and the fifth column showcases our final reconstruction. The last column demonstrates reconstructions from the recent BrainDiffuser work. All results are from subj01.

6.1 Ablation Study

To validate the contributions of our proposed extensions, we conducted ablation studies analyzing the impact of the depth estimation component. As shown in the attached table, we compared three model variations: 1) a baseline Stable Diffusion Img2Img pipeline using only the initial guess image, 2) a Depth2Image pipeline using only the estimated depth map, and 3) our full approach combining Stable Diffusion and ControlNet with both initial images and depth maps. Across low-level metrics like PixelCOrr and SSIM, the addition of depth information provided a consistent boost in performance. This aligns with the hypothesis that depth cues aid in capturing spatial relationships between objects and foreground-background segmentation. The full model with both initial images

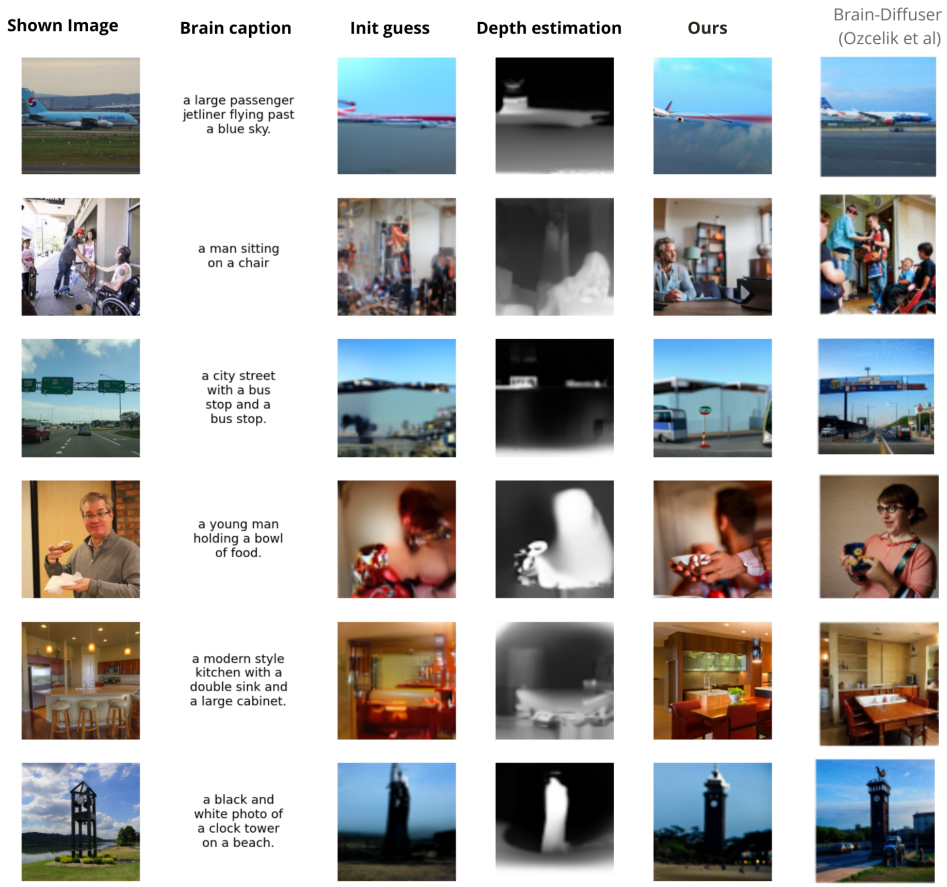

Figure A2: Comparison of our results (Columns 2-4) with the presented stimuli and other reconstruction works. The second column displays the caption derived from brain activity, the third column presents the initial guess image, the fourth column exhibits the depth-estimated images, and the fifth column showcases our final reconstruction. The last column demonstrates reconstructions from the recent BrainDiffuser work. All results are from subj01.

and depth performed the best, indicating that the two components are complementary. Qualitatively, the depth maps appeared to enhance object boundaries and 3D perspective. These results suggest that incorporating depth estimates helps the model reconstruct more accurate and realistic representations of the visual stimuli. The depth component specifically seems to benefit lower-level aspects like shapes and spatial relationships, which are critical for humans to perceive two images as highly similar Hermann et al. [2020]. By guiding the image reconstruction process with depth information extracted from brain activity, our approach can generate images that better match human perceptual judgments.

Ablation study	Low level metrics			High level		
Variant	PixCorr	SSIM	AlexNet (2)	AlexNet (5)	Inception	CLIP
Text + init	0.1204	0.1941	0.5815	0.7454	0.7974	0.8768
Stable Diffusion depth	0.3333	0.3106	0.8493	0.9654	0.8248	0.8778
ControlNet	0.**3379**	**0.3178**	**0.8707**	**0.9674**	**0.8238**	**0.8788**

Table 3: Ablation Study: Performance Metrics of Different Model Variants. Text + init is the plain Stable Diffusion Img2Img pipeline with initial guess image and captions predicted by the brain. Stable Diffusion depth is a variant pipeline that takes as input the initial guess image and captions and internally tries to estimate a depth map from the initial guess. ControlNet is external conditioning for the StableDiffusion Img2Img pipeline, so the inputs are the initial guess, the captions, and the depth maps estimated from the brain. This latter method is the one used in the paper and values (higher is better) show that this particular combination improves performance. Overall, this ablation study shows that including information about depth improves performances, particularly on low-level features.