Generative Multimodal Decoding: Reconstructing Images and Text from Human fMRI

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)

Abstract

The human brain adeptly processes immense visual information using complex neural mechanisms. Recent advances in functional MRI (fMRI) enable decoding this visual information from recorded brain activity patterns. In this work, we present an innovative approach for reconstructing meaningful images and captions directly from fMRI data, with a focus on brain captioning due to its enhanced flexibility over image decoding. We utilize the Natural Scenes fMRI dataset containing brain recordings from subjects viewing images. Our method leverages state-of-the-art image captioning and diffusion models for multimodal decoding. We train regression models between fMRI data and textual/visual features and incorporate depth estimation to guide image reconstruction. Our key innovation is a multimodal framework aligning neural and deep learning representations to generate both semantic captions and photorealistic images from brain activity. We demonstrate quantitative improvements in captioning over prior art and in image spatial relationships through our reconstruction pipeline. In conclusion, this work significantly advances brain decoding capabilities through an integrated vision-language approach. Our flexible decoding platform combining high-level semantic text and low-level visual depth information provides new insights into human visual cognition. The proposed methods could enable future applications in brain-computer interfaces, neuroscience, and AI.

1 Introduction

Brain decoding involves reconstructing sensory stimuli from recorded brain activity patterns. Here, we aim to decode visual information from fMRI recordings of brain activity during visual perception. We introduce a multimodal transformer model to generate natural language image captions directly from fMRI data. Predicting captions rather than images provides a higher-level interpretation of visual processing in the brain. Our model is trained on aligned fMRI-text-image triplets to learn

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

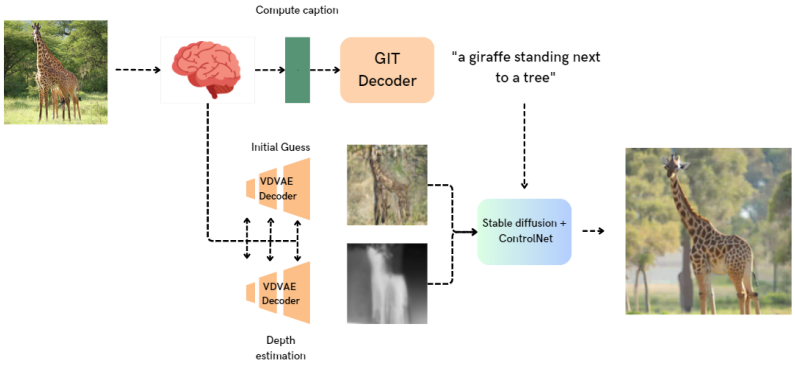


Figure 1: Our model utilizes fMRI measurements to extract features for GIT captioning and VDVAE initial and depth image estimation using linear models. Image captions serve as the primary general result, used in the second stage alongside other conditioning to generate plausible reconstructions with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

multimodal representations. We propose a conditioned image reconstruction pipeline leveraging predicted captions and depth maps from fMRI. The captions are used to control a state-of-the-art text-to-image diffusion model, Stable Diffusion, to generate photorealistic images conditioned on brain activity. Our main novelty is a multimodal framework exploiting alignments between fMRI, text, and images to decode both semantic captions and photorealistic images from brain activity. Figure 1 illustrates our full pipeline, while Figure 2 shows example generated captions and images compared to prior image reconstruction methods. We introduce novel multimodal decoding methods to predict higher-level textual scene descriptions and controlled photorealistic image reconstructions from fMRI recordings of visual perception. Our work provides new insights into visual information representation in the human brain.

2 Related Works

Several works have utilized variational autoencoders (VAEs) for decoding. VanRullen and Reddy [2019] employed a VAE-GAN to reconstruct faces, while Horikawa and Kamitani [2017] used sparse regression on fMRI data to predict CNN features. Other approaches have incorporated adversarial training. Shen et al. [2019], Ren et al. [2019], Gaziv et al. [2022] proposed unsupervised VAE-GANs for fMRI decoding. Donahue and Simonyan [2019], Casanova et al. [2021], Mozafari et al. [2020], Ozcelik et al. [2022] optimized latent spaces of GANs like BigBiGAN and IC-GAN. Recently, diffusion models have shown promise for improved image generation from fMRI [Takagi and Nishimoto, 2023, Chen et al., 2022, Ferrante et al., 2023, Ozcelik and VanRullen, 2023]. For caption generation, Takada et al., Matsuo et al. [2016], Qiao et al. [2018] combined CNNs and RNNs to estimate convolutional features and generate captions from brain activity. Our work proposes a neuroscience-inspired pipeline for multimodal brain decoding. We decode fMRI data into two processing streams: A top-down semantic stream predicts textual captions, mimicking higher-level semantic interpretation of visual scenes in the brain. A bottom-up stream decodes lower-level features like shapes, colors, and depth maps, emulating bottom-up visual processing. These dual streams are fused in a multimodal diffusion model aligning brain, text, and image representations. Our framework integrates neuroscience principles of hierarchical visual processing with state-of-the-art deep learning for multimodal decoding. The textual captions provide top-down conditioning, while the lower-level reconstructions inform bottom-up image details. In summary, we develop a neuro-inspired multimodal approach to reconstruct images by decoding both high-level semantic descriptions and low-level visual features from fMRI recordings of brain activity during visual perception.

3 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

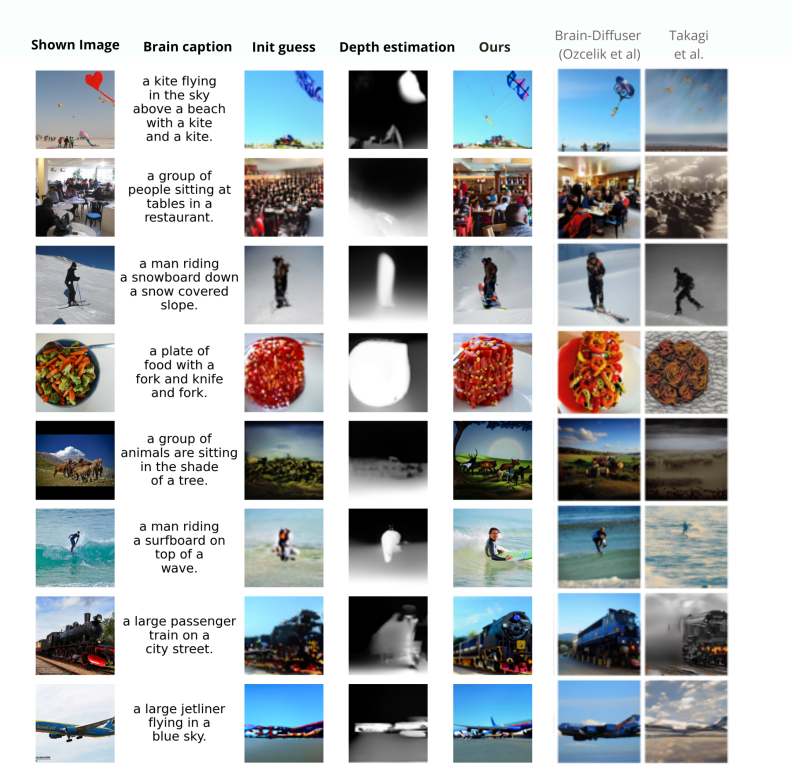


Figure 2: Comparison of our results (Columns 2-4) with the shown stimuli and reconstructions from other works. The second column displays the caption computed from the brain activity, the third column presents the initial guess image, the fourth column shows the depth estimated images, and the fifth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

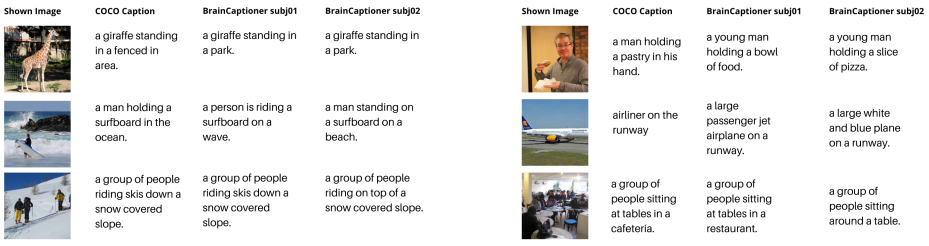


Figure 3: Examples of generated caption with our BrainCaptioner pipeline. Shown images are test set stimuli used for subj01 and subj02 during the fmri experiment. COCO Caption column report the first annotations for the original COCO image, while the other two columns are the output of our model for the two subjects.

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.

Data: We utilize the Natural Scenes fMRI dataset Allen et al. [2022] containing recordings from subjects viewing COCO images. Our analysis focuses on 4 subjects, yielding $\sim$ 25k train and $\sim$ 3k test trials. fMRI signals are masked and temporally reduced as in Allen et al. [2022].

Connecting Brain and deep learning models: Under the assumption that deep learning models can approximate the way the brain processes information, we can map one representation into another with a linear mapping. Specifically, we used a regularized linear regression between brain representations and latent space of an image captioning model to produce image captions from neural activity and a mapping to several layer representations of a variational autoencoder model to obtain initial image and depth reconstruction. Then, to improve image quality and combine these multiple

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**	**0.98**	0.84	0.90

Table 1: 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.

outputs, we used a multimodal latent diffusion model (Stable Diffusion + ControlNet) to generate the final image reconstruction.

Captioning: We employ the Generative Image-to-text Transformer (GIT) [Wang et al.] for captioning. We train ridge regression between fMRI data and GIT image features, and renormalize predictions to match the true distribution (Fig. A3). Generated captions are shown in Fig. 3.

Reconstruction: We propose a pipeline leveraging diffusion models conditioned on initial images and depth maps estimated from fMRI (Fig. 1). Initial images are generated by decoding VDVAE latents Child [2021] predicted from brain data. Depth is estimated using ControlNet Zhang and Agrawala [2023] to enhance foreground-background alignment. The full pipeline (Fig. 1) decodes fMRI to captions, initial images, and depth, then fuses these to reconstruct images via Stable Diffusion Rombach et al. [2021].

Evaluation: We compare captioning to a CNN-LSTM approach Takada et al., using metrics like METEOR and CLIP similarity. For image reconstruction, we evaluate on low-level high level image metrics reported in previous work including PixCorr, SSIM and 2-way accuracy in various CNN latent spaces.

4 Results

The table 1 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, A4, 3 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 3 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 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. Also, the alignment procedure between fMRI and text is based on human written short COCO caption descriptions. The use of dense captions with lot of details maybe generated using state-of-the art multimodal image captioning models could improve and stabilize the alignment procedure and outperform the current limitations.

5 Discussion and Conclusions

Applications: Our multimodal decoding approach has potential for neural art and style transfer by manipulating text prompts to guide image reconstruction from brain activity. This represents a new form of AI/neuroscience art. **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**: Key challenges include need for subject-specific models, insufficient high-quality training data, SNR constraints, reliance on simple mapping techniques, and involvement of multiple brain regions. Our captioning model also limits overall performance. **Conclusions**: We demonstrate promising results in multimodal caption and image reconstruction from brain activity. This work builds on neuroscience and AI concepts, furthering our understanding of visual and language processing in the brain. Refining this approach could lead to novel insights at the intersection of neuroscience and AI.

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.

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

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.

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.

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.

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.

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.

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.

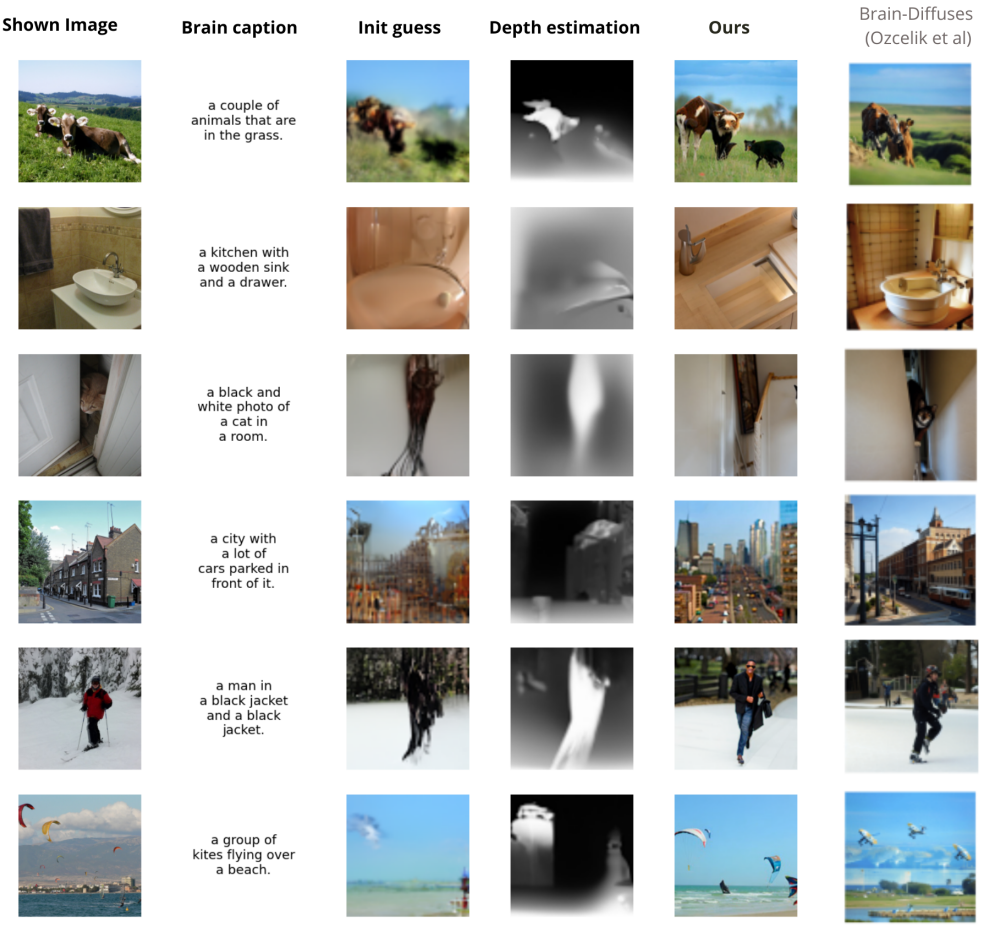


Figure A1: Comparison of our results (Columns 2-4) with the presented stimuli and other reconstruction works. The second column displays the caption derived from brain activity, the third column presents the initial guess image, the fourth column exhibits the depth-estimated images, and the fifth column 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

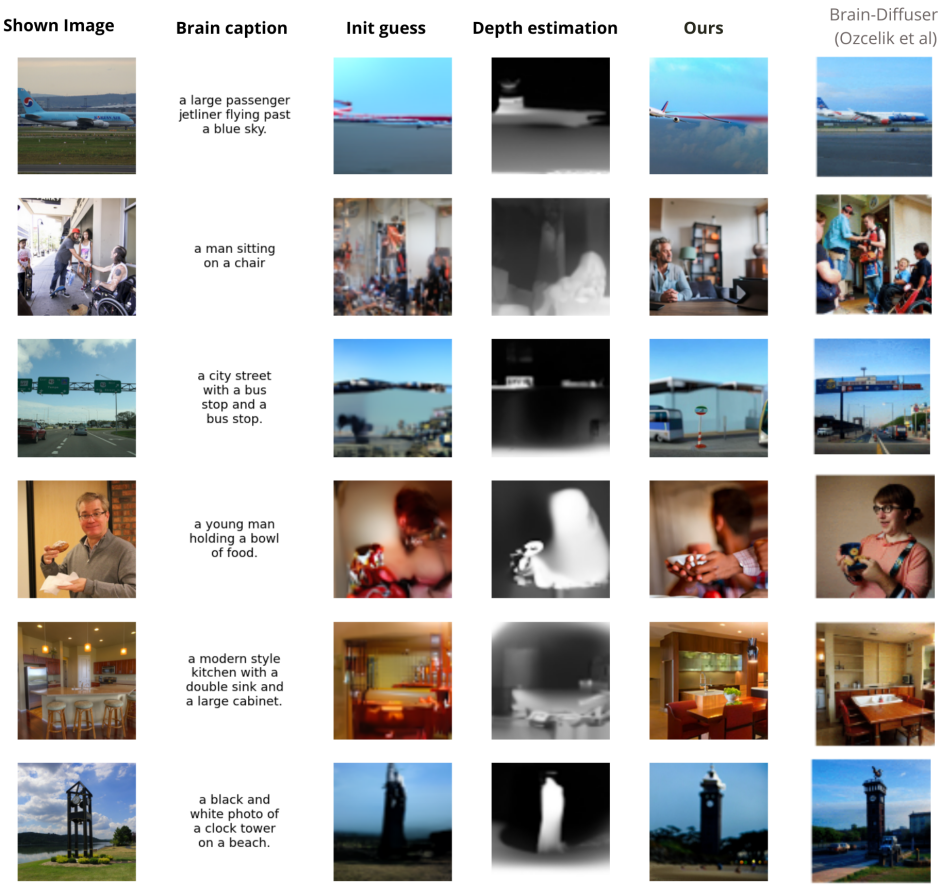


Figure A2: Comparison of our results (Columns 2-4) with the presented stimuli and other reconstruction works. The second column displays the caption derived from brain activity, the third column presents the initial guess image, the fourth column exhibits the depth-estimated images, and the fifth column 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.

Captioning Performances

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

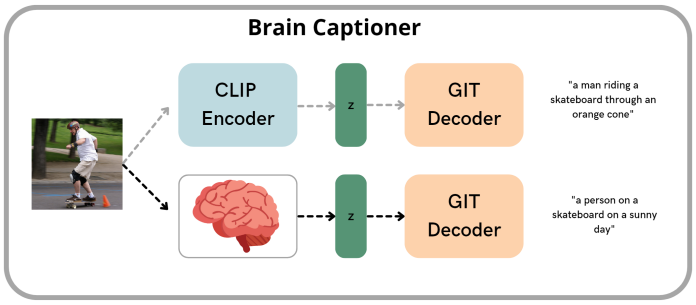
Ablation study	Low level metrics			High level		
Variant	PixCorr	SSIM	AlexNet (2)	AlexNet (5)	Inception	CLIP
Text + init	0.1204	0.1941	0.5815	0.7454	0.7974	0.8768
Stable Diffusion depth	0.3333	0.3106	0.8493	0.9654	0.8248	0.8778
ControlNet	0.**3379**	**0.3178**	**0.8707**	**0.9674**	**0.8238**	**0.8788**

Table 2: 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.

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 3: 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.

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.



1

Figure A3: 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.

Figure A4: 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.