Multi-Task Vision Experts for Brain Captioning

Anonymous CVPR submission

Paper ID 21

Abstract

Recent vision-language models have accelerated multimodal brain decoding, including brain captioning, which aims to translate brain activations into natural language. Typically, state-of-the-art models use either image or text for multimodal brain alignment. However, relying on image or text alone fails to capture the full range of brain activity. In this paper, we present MEVOX, a brain captioning method that introduces multi-expert vision systems for omni-contextual explanations. MEVOX ensembles multiple task-specific experts to capture distinct aspects of brain perception. By leveraging experts from various domains, we demonstrate that the proposed method can effectively pool this expert knowledge and adapt it to specific brain functions. Experimental results show that our method achieves competitive performance compared to current state-of-the-art methods.

1. Introduction

Recent advances in multimodal brain decoding have significantly deepened our understanding of how brain activations relate to visual perception and cognition [1]. These activations are captured when human participants view visual stimuli, and brain captioning aims to decode these neural recordings into natural language descriptions [6, 22, 28]. However, current brain captioning methods mostly rely on image or text as supervision for multimodal alignment, often failing to capture the full spectrum of brain activities and the intricate processing in neural functions.

The human visual system in the brain is known for its hierarchical structure, with each region specializing in specific functions [5, 9, 19, 29]. For example, V1 processes low-level features such as edge detection and orientation filtering. The organization of the visual brain follows a largely feed-forward structure, starting with V1, which receives direct input from the retina of the eye, and progressing to higher-level regions that represent increasingly abstract features. This hierarchical and specialized structure makes the brain a multi-tasking processor, analogous to corresponding computer vision tasks. Despite some exploration [28], this similarity has not been widely considered in brain decoding.

To bridge this gap, we propose using specialized vision experts for brain captioning to emulate the functional architecture of the visual cortex. By adopting a multi-expert perspective, we capture distinct aspects of brain perception, enhancing both the accuracy and contextual awareness of brain activity interpretations. This enables us to better replicate the hierarchical processing in the brain and provide more nuanced, task-specific insights into the underlying cognitive processes. Specifically, in this paper, we present MEVOX, a brain captioning method that introduces **M**ulti-**E**xpert **V**ision systems for **O**mni-contextual e**X**planations. Our method enlists a set of task-specific experts, each trained to capture particular facets of brain processing—ranging from low-level visual features to high-level semantic concepts. MEVOX mimics the multitasking within the visual cortex by pooling expert knowledge from diverse domains, enabling it to adapt to specific brain functions and capture the intricate, multi-faceted nature of brain activity. Experimental results demonstrate that MEVOX achieves competitive performance compared to state-of-the-art methods, offering a promising direction for future multimodal brain decoding research.

2. Related Work

Brain Captioning. Current brain captioning methods are broadly divided into two categories. The first category uses text annotations from COCO as supervision, learning a mapping (either through a trained encoder or direct regression) to predict text descriptions from brain signals using language models. For example, SDRecon employs ridge regression to align brain signals with intermediate representations of a language model, followed by a decoder to generate text descriptions. BrainCap [6] follows a similar pipeline but replaces the captioning model. OneLLM [7] introduces a unified encoder for multimodal-text alignment, improving caption quality but reducing CLIP similarity scores due to its exclusive alignment with text. The second category aligns brain signals with image features from the visual components of multimodal large language models (MLLMs), leveraging their inherent image captioning capabilities. This can be considered zero-shot since no text annotations are used. UM-

BRAE, for example, aligns brain signals with image features, preserving more accurate semantic and spatial information. The integration of LLMs further enhances the fluency, completeness, and informativeness of the generated captions.

Brain Alignment. The prevailing practice for multimodal brain alignment is to map neural modalities into a shared multimodal space [7, 17, 21, 28]. For example, OneLLM [7] employs generative training to learn multimodal alignment, connecting multimodal inputs, including brain signals, with an LLM. Other methods align brain signals within a pre-trained representation space, such as CLIP [20], using linear regression [17, 25], contrastive learning [29], diffusion priors [21], or feature reconstruction [28]. The target for alignment can be either embeddings or features, depending on the method. Since brain signals exhibit significant variations across subjects, most of these alignment strategies require per-subject training or subject-specific annotations, which can lead to scalability issues in practice. Therefore, another important aspect of brain alignment is mapping brain signals from different subjects to a shared representation space [22, 28], which can enhance performance and alleviate the challenges associated with individual subject variability.

3. MEVOX

This section details the method in Fig. 1. We first provide the rationale for connecting the human visual cortex with computer vision experts (Sec. 3.1). Next, we present the data preprocessing (Sec. 3.2) and the training objective (Sec. 3.3).

3.1. Human Visual Cortex

Studies have shown that the human brain follows a hierarchical structure, with each region specializing in specific functions. Different regions in the visual cortex process varying levels of visual granularity [5, 9, 19]. Specifically, early visual areas such as V1 and V2 encode low-level features like depth, edges, and shapes, corresponding to tasks such as depth estimation, normal estimation, and edge detection in computer vision. More complex semantic concepts are processed in areas like V4 and ITC, mirroring object detection and semantic segmentation processes. It is also well-established that different brain regions exhibit functional selectivity for categories such as faces, places, bodies, and words. The previous five tasks cover the first three categories from different perspectives but lack information on word processing. OCR (Optical Character Recognition) for text reasoning can be introduced to bridge this gap, enabling the model to capture brain functions related to word processing.

3.2. Vision Experts

We use six pretrained vision task experts as black-box predictors to produce multiple task-specific labels. These experts

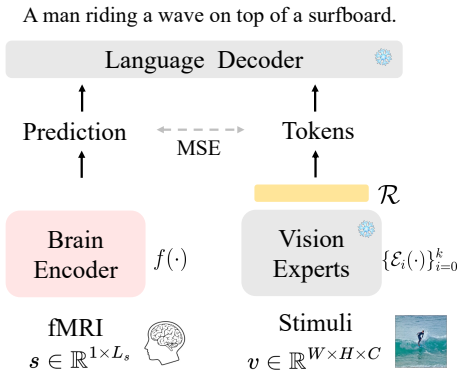

Figure 1. **Overview.** MEVOX ensembles multiple task-specific experts to capture distinct aspects of brain perception. The brain encoder learns to align with omni-contextual explanations.

encode three low-level vision signals (depth, surface normals, and edges) and three high-level vision signals (object labels, semantic labels, and OCR labels). Experts are frozen to retain domain-specific knowledge encoded in their network parameters. These expert models are relatively lightweight, incurring minimal additional training and inference costs with simple model parallelism. We apply task-specific post-processing on predicted labels from different vision experts, transforming them to a tensor $y_i \in \mathbb{R}^{H\times W\times C}$ (here H, W, C represent image height, width and channel respectively, *e.g.* $C = 1$ for depth and edge, and $C = 3$ for surface normal).

The most straightforward way is to use the labels from the six vision experts, with images as input, to train a corresponding model that decodes brain activations into specific domain labels. However, this requires training the same number of models as there are experts, making it inflexible and computationally intensive. Instead, we employ a multimodal feature adaptor [14] $\mathcal{R}$ to merge the predicted labels. The adaptor learns a predefined number of latent input queries to cross-attend to a flattened embedding concatenated from all multi-task features. It first maps the features to a uniform dimensionality and then processes a variable number of expert labels, outputting a fixed number of tokens. Specifically, a given image v is first fed into vision expert $\mathcal{E}_i$ to produce a set of predictions $\{y_i\}_{i=0}^{k} = \mathcal{E}_i(v)$, where each y_i represents the output of the i-th vision expert $\mathcal{E}_i$ and k is the total number of experts. The set of predicted features $\{y_i\}_{i=0}^{k}$ is then passed into the adaptor $\mathcal{R}$, which generates tokens c to produce the tokens that are fed into the language decoder $\mathcal{D}$ to produce the final captions. The full process from v to the tokens $c \in \mathbb{R}^{B\times N\times L}$ can be expressed as:

$$c = g(v) = \mathcal{R}\left(\{\mathcal{E}_i(v)\}_{i=0}^{k}\right) \quad \text{for } i = 0, 1, \ldots, k. \quad (1)$$

This ensures constant memory usage for self-attention in the vision encoder and vision-text cross-attention in the language decoder, independent of expert numbers.

Table 1. **Brain Captioning on BrainHub**. Models with '-S' refers to a cross-subject model trained in a single-subject setting. MindEye2 results are based on a model pre-trained on seven subjects and then adapted to the remaining one with the full dataset. The mark† denotes zero-shot methods, indicating no external captions were used for training. The **best** and second-best performance are highlighted.

Method	LLM	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE	CIDEr	SPICE	CLIP-S	RefCLIP-S
SDRecon [25]	BLIP [10]	36.21	17.11	7.72	3.43	10.03	25.13	13.83	5.02	61.07	66.36
OneLLM [7]	OneLLM [7]	47.04	26.97	15.49	9.51	13.55	35.05	22.99	6.26	54.80	61.28
BrainCap [6]	GIT [27]	55.96	36.21	22.70	14.51	16.68	40.69	41.30	9.06	64.31	69.90
MindEye2 [22]	GIT [27]	54.82	38.60	26.49	18.16	17.54	43.77	55.70	10.97	67.54	**73.73**
ViT3D [23]	LLaVA [13]	57.19	37.17	23.78	15.85	18.60	36.67	49.51	12.39	65.49	-
UMBRAE-S†	Shikra [4]	57.63	38.02	25.00	16.76	18.41	42.15	51.93	11.83	66.44	72.12
UMBRAE [28]†	Shikra [4]	**59.44**	**40.48**	27.66	19.03	**19.45**	43.71	**61.06**	**12.79**	**67.78**	73.54
MEVOX-S†	RoBERTa [15]	56.84	37.66	25.30	17.49	18.52	43.17	49.89	10.33	63.81	69.60
MEVOX†	RoBERTa [15]	58.56	40.36	**28.09**	**20.11**	19.20	**44.47**	54.37	11.05	64.23	70.36

3.3. Training Objective

Let B denotes the metric space of brain signals and V refers to images, with f and g as the fMRI encoder and visual experts, respectively. Given brain responses $s \in \mathbb{R}^{1\times L_s}$ and corresponding visual stimuli $v \in \mathbb{R}^{W\times H\times C}$, we train the brain encoder f to align brain tokens b with image tokens $\mathbf{c}$ extracted from the vision encoder g, aiming to approximate $f(s) \approx g(v)$. Toward the objective, the most straightforward alignment is element-wise feature reconstruction [28]:

$$\mathcal{L}_{\mathrm{R}} = \mathbb{E}\|f(v) - g(s))\|_2^2, \tag{2}$$

where f learns to map brain signals from space B to image space V. During training, only parameters of the brain encoder f are updated, while visual experts g and language decoders $\mathcal{D}$ remain fixed with pre-trained weight.

4. Experiments

4.1. Implementation Details

We trained a transformer-based encoder [28] to project brain voxel signals into the same embedding space as image features for alignment. We leveraged the pre-trained Prismer [14] as the vision expert encoder to produce the omni-contextual visual tokens $c \in \mathbb{R}^{B\times N\times L}$ as the supervisry signal, where B is the batch size, token number N is 1,220, and token dimension L is 1,024. All experiments were conducted on a single A100 GPU. We use the AdamW optimizer [16] with β_1=0.9, β_2=0.95, and a weight decay of 0.01. The learning rate was scheduled using the one-cycle strategy [24], starting with an initial learning rate of 3e-4.

4.2. Brain Captioning

We conducted all experiments on the Natural Scenes Dataset (NSD) [1], which contains fMRI signals recorded during experiments and visual stimuli from COCO [12]. Following prior studies [21, 25, 28, 29], we adopt the standard train and test splits for four subjects (subjects 1, 2, 5, 7), each with 24,980 training samples and a shared 982 test samples. For evaluation, we report the average of the three repetitions of the same image in the test set, totaling 982 samples per subject. For brain captioning, we use RoBERTa [15] as a frozen language decoder to generate natural language descriptions. For quantitative comparison, we adopt BrainHub [28], using ground truth captions sourced from COCO [12]. The generated captions are evaluated with seven standard metrics: BLEU-k [18], METEOR [3], ROUGE-L [11], CIDEr [26], SPICE [2], CLIP-Score [20], and RefCLIP-Score [8].

We train the fMRI encoder to align the voxel tokens with the omni-contextual visual tokens from the multi-task ensemble. The model training follows two scenarios: single-subject training (subject 1) and cross-subject training (subjects 1, 2, 5, and 7). In cross-subject training, two subjects are randomly selected in each epoch to compute the reconstruction loss for parameter updates. By learning a shared representation, the model aligns fMRI signals with visual tokens, enabling effective cross-modal correspondence. The language model will load the voxel tokens instead of image tokens to describe image details. Compared with prior studies [7, 25, 28], our method achieves state-of-the-art performance across metrics, except for the cross-subject UMBRAE [28]. The performance gap may be attributed to the use of more powerful MLLMs. MindEye2 [22] results are from a model pre-trained on seven subjects and adapted to the remaining one, with additional training subjects boosting performance. The experimental results demonstrate the effectiveness of introducing vision experts to pool omni-contextual knowledge and adapt it to specific brain functions.

5. Conclusion

In this paper, we present a novel brain captioning method that emulates the hierarchical structure of the visual cortex. Grounded in the principles of the human visual system, our method aligns brain activation with multi-expert vision systems for omni-contextual explanations. Our method, despite zero-shot, achieves competitive performance compared to the state-of-the-arts, demonstrating the effectiveness of integrating specialized vision experts for brain caption decoding.

References

[1] 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, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022. 1, 3

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, pages 382–398. Springer, 2016. 3

[3] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL Workshop*, pages 65–72, 2005. 3

[4] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 3

[5] Robert Desimone, Thomas D Albright, Charles G Gross, and Charles Bruce. Stimulus-selective properties of inferior temporal neurons in the macaque. *Journal of Neuroscience*, 4 (8):2051–2062, 1984. 1, 2

[6] Matteo Ferrante, Furkan Ozcelik, Tommaso Boccato, Rufin VanRullen, and Nicola Toschi. Brain captioning: Decoding human brain activity into images and text. *arXiv preprint arXiv:2305.11560*, 2023. 1, 3

[7] Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. Onellm: One framework to align all modalities with language. In *CVPR*, 2024. 1, 2, 3

[8] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP*, 2021. 3

[9] Nancy Kanwisher, Josh McDermott, and Marvin M Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of neuroscience*, 17 (11):4302–4311, 1997. 1, 2

[10] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900. PMLR, 2022. 3

[11] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 3

[12] 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 *ECCV*, pages 740–755, 2014. 3

[13] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 3

[14] Shikun Liu, Linxi Fan, Edward Johns, Zhiding Yu, Chaowei Xiao, and Anima Anandkumar. Prismer: A vision-language model with multi-task experts. *TMLR*, 2024. 2, 3

[15] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 3

[16] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 3

[17] Furkan Ozcelik and Rufin VanRullen. Brain-Diffuser: Natural scene reconstruction from fMRI signals using generative latent diffusion. *Scientific Reports*, 13(1):15666, 2023. 2

[18] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318, 2002. 3

[19] Aina Puce, Truett Allison, Maryam Asgari, John C Gore, and Gregory McCarthy. Differential sensitivity of human visual cortex to faces, letterstrings, and textures: a functional magnetic resonance imaging study. *Journal of neuroscience*, 16(16):5205–5215, 1996. 1, 2

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 2, 3

[21] Paul S Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Ethan Cohen, Aidan J Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, et al. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. In *NeurIPS*, 2023. 2, 3

[22] Paul S Scotti, Mihir Tripathy, Cesar Kadir Torrico Villanueva, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A Norman, and Tanishq Mathew Abraham. Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. In *ICML*, 2024. 1, 2, 3

[23] Guobin Shen, Dongcheng Zhao, Xiang He, Linghao Feng, Yiting Dong, Jihang Wang, Qian Zhang, and Yi Zeng. Neuro-vision to language: Enhancing brain recording-based visual reconstruction and language interaction. In *NeurIPS*, pages 98083–98110, 2024. 3

[24] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, pages 369–386. SPIE, 2019. 3

[25] Yu Takagi and Shinji Nishimoto. Improving visual image reconstruction from human brain activity using latent diffusion models via multiple decoded inputs. *arXiv preprint arXiv:2306.11536*, 2023. 2, 3

[26] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575, 2015. 3

[27] 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. *TMLR*, 2022. 3

[28] Weihao Xia, Raoul de Charette, Cengiz Oztireli, and Jing-Hao Xue. Umbrae: Unified multimodal brain decoding. In *ECCV*, pages 242–259, 2024. 1, 2, 3

[29] Weihao Xia, Raoul de Charette, Cengiz Öztireli, and Jing-Hao Xue. Dream: Visual decoding from reversing human visual system. In *WACV*, pages 8226–8235, 2024. 1, 2, 3