UMBRAE: Unified Multimodal Brain Decoding

Weihao Xia[1] Raoul de Charette[2] Cengiz Oztireli[3] Jing-Hao Xue[1]

[1]University College London [2]Inria [3]University of Cambridge

Abstract. We address prevailing challenges of the brain-powered research, departing from the observation that the literature hardly recover accurate spatial information and require subject-specific models. To address these challenges, we propose UMBRAE, a unified multimodal decoding of brain signals. First, to extract instance-level conceptual and spatial details from neural signals, we introduce an efficient universal brain encoder for multimodal-brain alignment and recover object descriptions at multiple levels of granularity from subsequent multimodal large language model (MLLM). Second, we introduce a cross-subject training strategy mapping subject-specific features to a common feature space. This allows a model to be trained on multiple subjects without extra resources, even yielding superior results compared to subject-specific models. Further, we demonstrate this supports weakly-supervised adaptation to new subjects, with only a fraction of the total training data. Experiments demonstrate that UMBRAE not only achieves superior results in the newly introduced tasks but also outperforms methods in well established tasks. To assess our method, we construct and share with the community a comprehensive brain understanding benchmark BrainHub. Our code and benchmark are available at `https://weihaox.github.io/UMBRAE`.

Keywords: Multimodal Brain Decoding · Universal Brain Encoder · Cross-Subject Training · Weakly-Supervised Adaptation

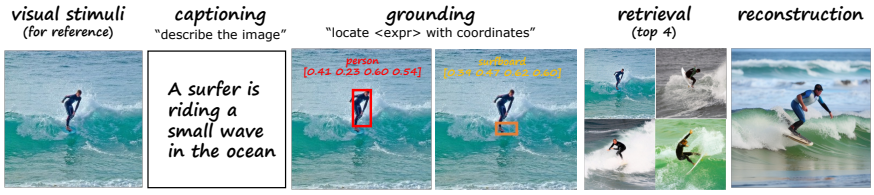

Fig. 1: Multimodal Decoding. By aligning brain features with MLLMs, UMBRAE decodes multimodal cues from brain signals, which allows multiple downstream tasks.

1 Introduction

Typically, artificial intelligence research relies on intermediate modalities to interpret human intentions, such as language [7, 44], gaze [1, 50], facial expression [10],

and action [16]. These modalities, however, are indirect channels of communication with humans and may be highly inaccurate for people with cognitive or physical disabilities or even locked-in patients who are conscious but unable to communicate through speech, limb, or facial movements [21]. In this context, the potential for direct interpretation of neural signals stands out as a promising prospect. The brain imaging literature has recently advanced, decoding neural signals into various forms such as image [37], video [8], or text [43] to read intentions [6, 22]. This deepens our understanding of the brain which neural activity is not directly comprehensible to humans.

However, there are remaining challenges in brain-powered research. First, decoding brain signals into a single modality results in a lossy representation of the brain activities. On the one hand, text fails to preserve the peculiar appearance of a texture or the spatial location of an object. On the other hand, visual decoding [31, 37, 49] addresses the underdetermined problem of pixel-wise reconstruction and lacks explicitation of the scene structure. Consider, for instance, the scenario where a person uses thoughts to control a robotic arm to retrieve an apple from a fruit bowl on a table. The first task is to recognize the apple amidst similar visual concepts and then locate its exact position. But current methods lack such fine-grained decoding capability to interpret object categories, visual concepts, and their relationships. The second challenge pertains to the subject-specific patterns of brain activities [2]. Therefore, current methods typically train a model for each subject to cope with distinctive brain patterns. Decoding brain signals across multiple subjects presents challenges due to the structural and functional differences among individual brains.

Hence, we instead propose to decode a robust multimodal representation which serves as proxy for downstream tasks, such as textual or visual decoding. Our method allows brain decoding at different granularities, through prompting, which unravels unprecedented brain-machine interface for locked-in patients [21] that typically requires iterative feedback. To evaluate our novel tasks, we extend the popular Natural Scenes Dataset (NSD) [2] with multimodal ground truth, which constitutes a new brain understanding benchmark. Both code and benchmark will be made publicly available. Our contributions summarize as follows:

- We introduce UMBRAE, aiming at unified multimodal brain decoding. Our method relies on a universal brain encoder and a frozen multimodal large language model seeking to align brain signals with images. We also propose a cross-subject training strategy to learn a universal representation across subjects, as opposed to the standard subject-specific training. Furthermore, it allows the novel weakly-supervised adaptation, enabling the training of a model for any arbitrary subject with minimal training data.
- We construct BrainHub, a multimodal brain understanding benchmark extending NSD [2]. The benchmark pairs fMRI with semantic concepts and spatial localization in visual stimuli, offering tasks and metrics for evaluation.
- Our method achieves better or on par performance compared to state-of-the-art methods on a variety of tasks including brain captioning, retrieval,

and visual decoding. It is also the first one to enable direct brain grounding, performing on par with natural baselines while being at least 10 times faster.

2 Related Works

Brain-Conditioned Generation. Generative vision models conditioned on brain signals [25, 31, 37, 40, 49] have recently achieved unparalleled performance in decoding visual stimuli from corresponding brain responses. Generally, these methods map brain responses, captured in the form of functional magnetic resonance imaging (fMRI), to more common modalities suitable for feeding into pretrained vision-language models [19, 36, 51] for subsequent image reconstruction. For example, Lin *et al*. [25] project fMRI data to a CLIP [35] common space and reconstruct images through a finetuned StyleGAN2 [19]. Takagi and Nishimoto [40] utilize the ridge regression to link fMRI signals with CLIP text embeddings and the latent space of Stable Diffusion [36] (SD). Xia *et al*. [49] extract semantics, depth, and color cues, and reconstruct images using a depth-color-conditioned SD. Rather than relying solely on textual embeddings, several methods [14, 41] aim to obtain explicit descriptions for the visual stimuli. In contrast, our method decodes brain responses into various human-readable textual and visual cues, which can also flexibly serve as inputs for generative models.

Multimodal Large Language Models. Expanding Large Language Models (LLMs) to encompass other modalities, such as images, has garnered considerable attention recently. These models typically comprise three components: a frozen image encoder, a trainable adapter, and a frozen or finetuned LLM. The adapter's role is to bridge the gap from image features to the LLM, which can be implemented as a linear layer [7], a multilayer perceptron (MLP) [27], or a lightweight transformer [17]. In addition to vision-focused LLMs, recent studies aim to expand the boundaries of LLMs to include other modalities, making it possible to unify multiple modalities within a single LLM. Brain signals, as an emerging modality, have also recently been incorporated, for example in OneLLM [14], but like training with other modalities, all these methods require massive amounts of data and abundant computational resources. In contrast, we demonstrate a simple yet effective way to align brain signals with images.

Multimodal-Brain Alignment. The prevailing practice for brain alignment is to map the neural modality into a common latent space [14, 31, 37, 49], which can be divided into two lines of works: discriminative alignment and generative alignment. Considering the scarcity of data, methods in the first category aligns the brain modality within a pretrained embedding space, such as CLIP [35], through direct regression [31,40,41], contrastive learning [49], or diffusion prior [37]. The second is garnering significant attention in the field of MLLMs. For instance, OneLLM [14] adopts generative training to learn the alignment of multimodal inputs, including brain signals, thereby connecting a universal encoder with an LLM. However, such alignments between brain signals and images or text are trained per subject, resulting in one model for each subject. In contrast, we align

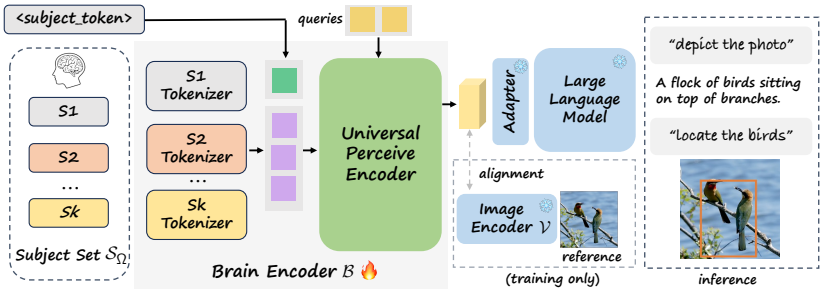

Fig. 2: Overview of UMBRAE. Our brain encoder includes subject-specific tokenizers and a universal perceive encoder (Sec. 3.1). Neural signals (fMRI) from multiple subjects are mapped into a common feature space, enabling cross-subject training and weakly-supervised adaptation (Sec. 3.2). The brain encoder learns to align neural signals with image features (Sec. 3.3). During inference, the learned encoder interacts with MLLMs and performs brain understanding tasks according to given prompts (Sec. 3.4).

the brain modality with image features to recover both semantic and spatial cues and achieve cross-subject multimodal-brain alignment to leverage user diversity.

3 UMBRAE

Our method is designed to address two shortcomings of the brain decoding literature. First, instead of learning a unimodal decoding (text or image), we learn a brain encoder that aligns brain features with pretrained multimodal space thus benefiting from all the MLLMs downstream tasks for multimodal decoding. Second, we observe that prior works train per-subject models owing to neuroscience showing inter-subject variability of the brain activities [2]. Instead, we intuit that inter-subject patterns present and learn a subject-unified representation by training jointly across-subjects. Beside better performance, this allows adaptation to novel subjects with minimal training data.

The overview of our method, named UMBRAE, is in Fig. 2. The acronym stands for unified multimodal brain decoding and signifies the process of unveiling encoded information hidden within the 'shadows' of brain signals. We rely mainly on a flexible brain encoder architecture (Sec. 3.1) and a cross-subject training strategy (Sec. 3.2) to map brain responses from different subjects, each with variable length, into a common feature space. In our experiments, we observe that simply binding brain modality with image features enables the recovery of both semantic and spatial cues (Sec. 3.3). We then conduct brain prompting interface by inputting the standardized brain representations from diverse subjects into MLLMs for downstream tasks (Sec. 3.4).

3.1 Architecture

Our brain encoder, based on a lightweight transformer architecture [17, 46], accommodates variable-length brain response inputs. This is important for cross-

subject training as fMRI data are variable across subjects. Hence, our architecture comprises subject-specific tokenizers which aim is to extract and map the subject-specific characteristics, along with a universal perceive encoder designed to capture subject-agnostic knowledge, later aligned with image features.

Subject Tokenizer. The subject-specific tokenizer projects the input brain signal $s \in \mathbb{R}^{1\times L_s}$ selected from the subject set $\mathcal{S}_\Omega$, with arbitrary length L_s, into a fixed-length sequence of brain tokens $\mathbf{x} \in \mathbb{R}^{L\times D}$. Here, L is the sequence length and D is the token dimension. Considering inter-subject variability in brain patterns [2], we design a separate tokenizer for each subject. Besides, we introduce a learnable subject-specific token $\{\mathbf{s}_k\}_{k=1}^{K}$ to switch between subjects, where K is the total number of subjects and $\mathbf{s}_k \in \mathbb{R}^{M\times D}$ contains M tokens of dimension D. Then, we prepend subject-specific tokens $\mathbf{s}_k$ to the predicted brain tokens $\mathbf{x}$ and encode them with the following universal perceive encoder.

Universal Perceive Encoder. The universal perceive encoder seeks to project all brain tokens $\mathbf{x}$ into a common space. We utilize here a transformer-based architecture [17] which uses cross-attention modules to project the input tokens into a latent bottleneck where the key K and value V are projections of the input tokens, while Q is the projection of learnable latent queries.

The subject-specific tokenizers are expected to capture specific information for each subject, including structural and functional differences among individual brains; and the universal perceive encoder aims to extract common knowledge across different subjects, encompassing categories, semantics, textures, and geometries of various objects and scenes. We now detail the training strategy for cross-subject alignment.

3.2 Cross-Subject Alignment

For cross-subject alignment, it is crucial to ensure that examples from each subject are uniformly sampled. This enables the model to avoid subject preference and prevent catastrophic forgetting. Therefore, we adopt a sampling strategy to ensure that θ percent of samples in a batch are from the same subject. Considering $\mathcal{S}_\Omega$ being the union over K subjects training data $\mathcal{S}_\Omega = \bigcup_{k\in\{1,2,\dots,K\}} \mathcal{S}_k$, we select data samples from a subject $\mathcal{S}_k$ with probability:

$$p_k = \frac{\|\mathcal{S}_k\|}{\sum_{n=1}^{K} \|\mathcal{S}_n\|}, \tag{1}$$

where $\|\cdot\|$ denotes cardinality (*i.e.*, the number of data samples). To construct a batch with B number of data samples, we select a subject S_k with probability p_k and conduct random sampling to yield $\theta \times B$ training samples. The remaining $(1-\theta)\times B$ examples are uniformly sampled from other subjects, *i.e.*, $\mathcal{S}_\Omega \backslash \mathcal{S}_k$. This batch sampling strategy significantly benefits from user diversity as it allows the model to focus primarily on intra-subject training while being exposed to different subjects to improve inter-subject discrimination. We latter demonstrate that this cross-subject alignment enhances performance without incurring extra training costs compared to training with a single subject, at no additional training time.

3.3 Multimodal Alignment

Previous multimodal methods learn to map multiple modalities into a common latent space, typically through contrastive pretraining, using either images [13] or text [14] as the binding modality. Instead, we align brain representations with image features from a pretrained image encoder using element-wise reconstruction.

Given a brain response $s \in \mathbb{R}^{1\times L_s}$ and the corresponding visual stimulus $v \in \mathbb{R}^{W\times H\times C}$, the source brain encoder $\mathcal{B}$ and target visual image encoder $\mathcal{V}$ encode brain signals and images into features denoted $\mathbf{b}$ and $\mathbf{v}$, respectively. We train the brain encoder $\mathcal{B}$ to minimize the distance between brain features and image features, aiming for a close approximation $\mathcal{B}(b) \approx \mathcal{V}(v)$ through:

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_{\mathbf{b}\sim\mathbf{B},\mathbf{v}\sim\mathbf{V}}[\|\mathcal{V}(v) - \mathcal{B}(b))\|_2^2] \,. \tag{2}$$

Our brain encoder $\mathcal{B}$ learns the alignment between source brain space $\mathbf{B}$ and target image space $\mathbf{V}$. Different from previous methods, we align the brain signals with the intermediate image features from a pretrained image encoder, thus achieving semantic and spatial alignment for the brain representation. MLLMs [7, 27] show such features provide sufficient visual cues for finetuning LMMs. Furthermore, aligning with intermediate image features allows direct input of aligned brain representations into the MLLM.

3.4 Brain Prompting Interface

After alignment, brain features from the brain encoder $\mathcal{B}$ are fed into the MLLM's adapter, to retrieve the mapped visual embeddings `<image>`. These embeddings are then concatenated with a user instruction prompt and inputted into the finetuned LLM. Thus, our brain encoder inherits from multimodal capabilities of the MLLM, allowing tasks to be used in a prompting fashion using template:

> **system message.** user: <instruction> **<image>** assistant: <answer>

The tags <instruction> and <answer> serve as placeholders for human instructions and assistant answers. We use variable templates for different tasks. Specifically, brain captioning uses '**Describe this image <image> as simply as possible.**'; brain grounding, '**Locate <expr> in <image> and provide its coordinates, please.**', where **<expr>** is the expression. More templates for different supported tasks can be found in the supplementary material.

Our primary focus in this study is on brain captioning and grounding, which reflects the capabilities of brain signals in concept recognition and spatial localization. They are often referred to as image captioning and visual grounding in the multimodal learning literature. However, in this context, brain signals are used as the input rather than images. Our method also supports other instruction-following capabilities, such as conversation, detail description, and complex reasoning. Our method is model-agnostic, allowing for the use of any image encoders, LLMs, and MLLMs according to specific needs.

4 Experiments

4.1 Implementation Details

Architecture. We use the pretrained CLIP ViT-L/14 [35] as the visual encoder and Vicuna-7B/13B [9] as the LLM, consistent with the setup in Shikra [7] and LLaVA [27]. The target image features are obtained from the second last layer of the transformer encoder, denoted as $\mathbf{T} \in \mathbb{R}^{16\times16\times1024}$, which are then converted to $\mathbf{T}' \in \mathbb{R}^{256\times D}$ for further processing by the adapter and LLM. The dimension D is 4,096 for Vicuna-7B and 5,120 for Vicuna-13B. The learnable tokens for each subject are of dimensions $\mathbb{R}^{5\times1024}$.

Training Details. Our models are trained on a single A100 GPU for 240 epochs with a global batch size of 256. It takes around 12 hours to complete. We use AdamW [28] as the optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and weight decay of 0.01. For the learning rate scheduler, we use one-cycle [38] with an initial learning rate of 3e-4. We set $\theta = 0.5$, meaning that in each batch of 256 samples, 128 come from each of two subjects. The selection probabilities are identical for each subject, as they contain the same number of training data. Following visual decoding studies [37, 40, 49], we use the standard train and test splits for the four subjects (S1, S2, S5, S7). Specifically, each subject contains 24,980 training samples. For testing, we report the average of the three same-image repetitions, totaling 982 samples per subject. Note that the above studies train a subject-specific model for each of the four subjects, while we train one brain encoder for them all.

4.2 BrainHub

For evaluation, we construct a multimodal brain understanding benchmark, BrainHub, to further analyze the information contained in brain signals. Specifically, we extend the NSD [2], a popular dataset comprising brain responses of subjects viewing visual stimuli (images) sourced from Microsoft Common Objects in Context (COCO) [26]. NSD provides (fMRI, image) pairs which is sufficient for visual decoding. However, we aim to explore the ability to process brain signals for identifying visual concepts, recognizing and localizing instances, as well as extracting spatial relationships among multiple exemplars. Specifically, we process the corresponding COCO images for each fMRI sample and extract relevant labels for the following tasks and metrics:

- **Brain Captioning** aims at textually describing the primary content of a given brain response. Ground truth captions are retrieved from COCO [26], and evaluation of inferred captions uses five standard metrics: BLEU-k [32], METEOR [4], ROUGE-L [24], CIDEr [47], and SPICE [3], in addition to two CLIP-based scores [35], namely CLIP-S and RefCLIP-S [15].
- **Brain Grounding** is the counterpart of visual grounding [7, 27] and seeks to recover spatial locations from brain signals by inferring coordinates. Given that identified classes might be named differently, or simply absent from

Table 1: Brain Captioning. 'UMBRAE-S1' refers to our model trained with a single subject (S1 here) only, while 'UMBRAE' denotes the model with cross-subject training. 'Shikra-w/img' refers to the image captioning result from Shikra [7] using the ground truth image as input, serving as an approximate upper bound. The colors represent the best, second-best, and third-best performance.

Method	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE	CIDEr	SPICE	CLIP-S	RefCLIP-S
Shikra-w/img [7]	82.38	69.90	58.63	49.66	35.60	65.49	161.43	27.62	80.60	85.92
SDRecon [40]	36.21	17.11	7.72	3.43	10.03	25.13	13.83	5.02	61.07	66.36
OneLLM [14]	47.04	26.97	15.49	9.51	13.55	35.05	22.99	6.26	54.80	61.28
UniBrain [29]	-	-	-	-	16.90	22.20	-	-	-	-
BrainCap [12]	55.96	36.21	22.70	14.51	16.68	40.69	41.30	9.06	64.31	69.90
UMBRAE-S1	57.63	38.02	25.00	16.76	18.41	42.15	51.93	11.83	66.44	72.12
UMBRAE	59.44	40.48	27.66	19.03	19.45	43.71	61.06	12.79	67.78	73.54

ground truth labels, we evaluate bounding boxes through the task of referring expression comprehension [52], using accuracy and intersection over union (IoU) as the evaluation metrics.

- **Brain Retrieval** is to search for pertinent results in response to a provided query from a large database, often considered as a form of fine-grained, instance-level classification. The evaluation metric used is accuracy.
- **Visual Decoding** refers to the capability to reconstruct the visual stimuli associated with the fMRI data. We include it here for consistency with the extensive literature on visual decoding [31, 37].

4.3 Brain Captioning

Tab. 1 provides an evaluation of our brain captioning for subject 1 (S1), with respect to SOTA baselines being SDRecon [41], BrainCap [12] and OneLLM [14]. From the latter table, UMBRAE outperforms all baselines by a significant margin on all metrics. SDRecon poor performance results from its limited limited vocabulary, and the use redundant or meaningless words in its captioning, such as '*person and person with person person wearing a tie shirt person person, women's clothing.*', which impacts the quality metrics negatively. BrainCap [12] follows a similar pipeline but replaces the captioning model, which performs better. OneLLM [14] learns a unified encoder for multimodal-text alignment which improves the caption quality but deteriorates the CLIP similarity score, as it merely aligns with texts. In contrast, the alignment with image features of UMBRAE preserves more accurate semantic and spatial cues decoded from the brain signals. Moreover, the use of LLMs helps generate sentences that are fluent, complete, and rich in information. Interestingly, we note that the performance of UMBRAE (trained on S1, S2, S5, S7) exceeds those when trained only on data from subject 1 (UMBRAE-S1), demonstrating the ability to learn from cross-subject patterns. As a approximate upper bound, we also report 'Shikra-w/img' which, similar to us, utilizes Shikra [7] for captioning though here using the ground truth image (visual stimuli). Results for other subjects are provided in the supplementary material.

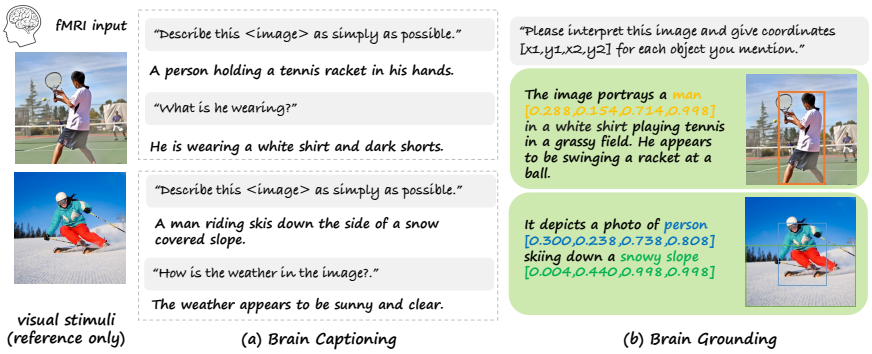

Fig. 3: Example Results. Our method inherits the multimodal capability from MLLMs and thus supports multiple brain captioning and grounding tasks. Different task prompts for the same input brain signal result in unique outcomes.

4.4 Brain Grounding

Our method showcases grounding capabilities across various settings by adapting to the corresponding instructions, which is illustrated in Fig. 3. For instance, we conduct spotting captioning, a task aimed at generating a description of the image along with bounding boxes for the mentioned items, using the instruction '`Please interpret this image and provide coordinates [x1,y1,x2,y2] for each object you mention`'. We can also perform referring expression comprehension using '`could you find and tell me the coordinates of <expr>?`'. For evaluation, we detect queried objects and report the accuracy and IoU. The accuracy metric 'acc@m' measures the percentage of correctly labeled instances with an IoU greater than the threshold m. Results of acc@0.5 are reported in Tab. 2 and examples are depicted in Fig. 3. More results are in the supplementary material.

In Tab. 2 we again report an approximate upper bound 'Shikra-w/img' being the *visual* grounding using the ground truth image. Given the absence of prior brain grounding baselines, we construct natural baselines by combining Shikra with the images from SOTA visual decoding methods [31, 37, 49], referred as 'Shikra-w/method'. We also report 'Shikra-w/UMBRAE' using our own visual decoding later described in Sec. 4.6. Being the first to attempt decoding spatial information from brain signals, our method 'UMBRAE' performs roughly on par with our constructed baselines while being at least x10 faster for grounding noting that speed is a critical characteristics for brain controlled applications.

In addition to metrics for all classes denoted 'All', we inspire from neuroscience exploring the salience-processing systems in the human brain [45] for more detailed evaluation. Specifically, we group the 80 classes of COCO [26] according to their prominence into: 'Salient', being the union of 'Salient Creatures' (people and animals) and 'Salient Objects' (*e.g.*, car, bed, table), and 'Inconspicuous' (*e.g.*, backpack, knife, toothbrush). We report the detailed mapping in the supplementary. It is interesting to note that UMBRAE is outperformed on Salient Creatures but not on Salient Objects and Inconspicious elements. This

Table 2: Brain Grounding. 'UMBRAE-Sx' refers to our model trained with a single subject only, while 'UMBRAE' denotes the model with cross-subject training. 'Shikra-w/img' refers to the *visual* grounding result from Shikra [7] using the ground truth visual stimuli (images) as input. Similarly, 'Shikra-w/method' provides visual grounding results using images produced by visual decoding methods [31, 37, 49]. We highlight best, second-best, and third-best performance per subject.

Method	Eval	All		Salient		Salient Creatures		Salient Objects		Inconspicuous		Time (s)
		acc@0.5	IoU	acc@0.5	IoU	acc@0.5	IoU	acc@0.5	IoU	acc@0.5	IoU	
Shikra-w/img [7]	*	51.96	47.22	62.92	56.44	66.71	59.34	58.79	53.27	38.29	35.71	0.96
Shikra-w/BrainDiffuser [31]	S1	17.49	19.34	27.18	27.46	38.71	34.63	14.62	19.66	5.39	9.20	16.4
Shikra-w/MindEye [37]		15.34	18.65	23.83	26.96	29.29	31.64	17.88	21.86	4.74	8.28	16.4
Shikra-w/DREAM [49]		16.21	18.65	26.51	27.35	34.43	33.85	17.88	20.28	3.35	7.78	10.5
Shikra-w/UMBRAE		16.83	18.69	27.10	27.55	34.14	33.65	19.44	20.92	4.00	7.64	16.4
UMBRAE-S1		13.72	17.56	21.52	25.14	26.00	29.06	16.64	20.88	4.00	8.08	0.92
UMBRAE		18.93	21.28	30.23	30.18	39.57	36.64	20.06	23.14	4.83	10.18	0.92
UMBRAE-S2	S2	15.21	18.68	23.60	26.59	27.86	30.51	18.97	22.32	4.74	8.81	-
UMBRAE		18.27	20.77	28.22	29.19	38.29	36.13	17.26	21.63	5.86	10.25	-
UMBRAE-S5	S5	14.72	18.45	22.93	26.34	26.86	29.84	18.66	22.52	4.46	8.60	-
UMBRAE		18.19	20.85	28.74	30.02	36.71	36.25	20.06	23.23	5.02	9.41	-
UMBRAE-S7	S7	13.60	17.83	21.07	25.19	24.57	28.90	17.26	21.15	4.28	8.64	-
UMBRAE		16.74	19.63	25.69	27.90	33.14	33.42	17.57	21.89	5.58	9.31	-

* The subjects test sets use the same reference images making 'Shikra-w/img' identical for all subjects.

suggests that visual decoding effectively reconstruct the salient creatures in the image space, arguably because the subject focuses on the latter.

Experimentally, we also notice that images containing few salient objects exhibit better performance compared to cluttered scenes, and easy background also lead to better grounding. Conversely, we note that localization suffers when images are filled with numerous inconspicuous objects. We argue that inconspicuous objects in the image may not draw the subject's attention, or that relevant brain activities may not be effectively captured during experiments [2]. Our categorization and observation also align with the semantic selectivity found in the higher visual cortex of the human brain [11, 18, 34], which contains specialization of certain regions that respond selectively to specific semantic categories of visual stimuli, such as faces, bodies, words, food, and places. The results demonstrate that our method performs well in relevant cases.

4.5 Brain Retrieval

The retrieval evaluation demonstrates the amount of image-specific information contained in the brain embedding. Following [37], we conduct three experiments: forward retrieval, backward retrieval, and exemplar retrieval. The *forward* retrieval computes accuracy of identifying the correct paired CLIP image embedding from 300 brain embeddings. Conversely, the *backward* retrieval finding the correct brain embedding from 300 image embeddings. For a fair comparison, we modify the output dimension and proceed to optimize the encoder and embedding using an InfoNCE [30] loss. We follow the same procedure as in [25] for calculating the retrieval metrics reported in Tab. 3. The *exemplar* retrieval aims to find the exact original image within the 982 test images. Our method outperforms

Table 3: Brain Retrieval. We report *forward*, *backward*, and *exemplar* retrieval metrics [37], showing that our method significantly outperforms the baselines. We also compare the floating-point operations (FLOPs), multiply-accumulate operations (MACs), and model parameters (Params). 'UMBRAE' denotes the model with cross-subject training. Colors represent the best and second-best performance.

Method	Forward	Backward	Exemplar	FLOPs (G)	MACs (G)	Params (M)
MindReader [25]	11.0%	49.0%	\	\	\	\
BrainDiffuser [31]	21.1%	30.3%	\	\	\	\
MindEye [37]	93.6%	90.1%	93.2%	52.27	26.13	1,003.64
UMBRAE	94.2%	91.3%	93.8%	67.48	33.72	146.24

Table 4: Visual Decoding. Following the standard evaluation metrics [31], our method performs on par or better than the SOTA methods [25, 31, 37, 40, 49]. Colors represent the best, second-best, and third-best performance.

Method	Low-Level				High-Level			
	PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	EffNet-B ↓	SwAV ↓
MindReader [25]	-	-	-	-	78.2%	-	-	-
SDRecon [40]	-	-	83.0%	83.0%	76.0%	77.0%	-	-
BrainDiffuser [31]	.254	.356	94.2%	96.2%	87.2%	91.5%	.775	.423
MindEye [37]	.309	.323	94.7%	97.8%	93.8%	94.1%	.645	.367
DREAM [49]	.274	.328	93.9%	96.7%	93.4%	94.1%	.645	.418
UMBRAE	.283	.341	95.5%	97.0%	91.7%	93.5%	.700	.393

current methods with accuracy percentages of 94.2%, 91.3%, and 93.8% on forward, backward, and exemplar retrieval experiments, respectively. These results demonstrate the ability to distinguish among misconstruable exemplars and suggest the fine-grained, image-specific information retained in the predicted brain embeddings.

4.6 Visual Decoding

Although this is not our primary purpose, to show the versatile capabilities of our method, we conduct experiments on the visual decoding task and compare with SOTAs [25, 31, 37, 40] using recognized metrics. While our method is not specifically tailored for this task, the textual and visual outputs it generates can be used as cues for fMRI-to-image reconstruction. Our results in Tab. 4 is utilizing the Versatile Diffusion [51] to reconstruct the image based on the decoded text and CLIP image embedding obtained in Sec. 4.5. We employ the same evaluation metrics as used in [31]. Specifically, PixCorr calculates the pixel-level correlation between the ground-truth and reconstructed images. SSIM [48] measures the textural and structural similarity instead of pixel-wise differences. Two-way comparisons are conducted using the second and fifth layers of AlexNet [20], the last pooling layer of Inceptionv3 [39], and the last layer of CLIP ViT-L/14 image encoder [35]. EffNet-B and SwAV are distances from EfficientNet [42] and SwAV-ResNet50 [5]. The first four metrics focus on low-level characteristics, whereas the remaining metrics are concerned with higher-level measurements.

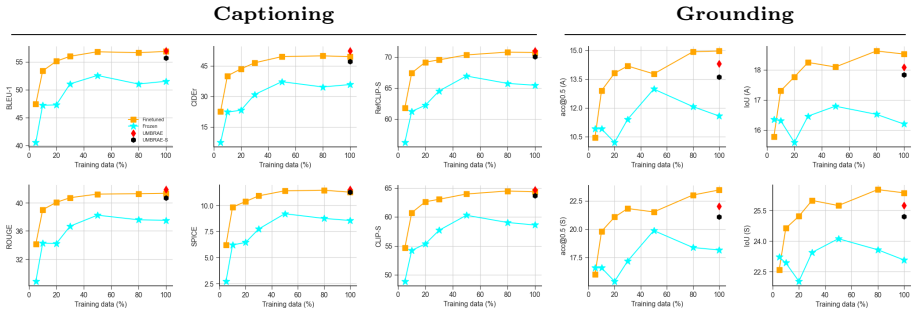

Fig. 4: Weakly-Supervised Subject Adaptation. This model for S7 is trained or finetuned on a pretrained model (trained on S1, S2, and S5) using varying ratios (0.05, 0.1, 0.2, 0.3, 0.5, 0.8, 1.0) of training data. The model achieves comparable performance using only 30% of the data and obtains better results when increasing the ratio of used training samples to 50%, compared to the model trained on the full dataset of S7.

Results in Tab. 4 demonstrate that our method performs comparably or better than state-of-the-art methods without any specific designs tailored for this task. Moreover, with access to common conditions like texts, image embeddings, and bounding boxes, we can leverage a wide range of pretrained image generation models. These models encompass text-to-image (*e.g.*, SD [36], SD-XL [33]), layout-to-image [23], and multiple-condition [51]. Details can be found in the supplementary material.

4.7 Weakly-Supervised Adaptation

Capturing brain signals, such as high-resolution fMRI, requires specialized equipment and professional personnel, making it challenging to collect on a large scale. A benefit of our cross-subjects training is to allow subject adaptation with minimal training data. To evaluate this emerging property, we train our brain encoder with subjects S1, S2, and S5 and seek to adapt the trained to a new subject S7 using various amount of training data. For ablation, we explore two settings where we train a new tokenizer for S7 with the universal perceive encoder being either *Frozen* or *Finetuned*.

Plots in Fig. 4 report 'Frozen' and 'Finetuned' adaptation with variable amount of S7 data. Additionally, we report 'UMBRAE' when trained with all training data of {S1, S2, S5, S7} as as well as 'UMBRAE-S7' trained on all S7 data only. Notably, compared to 'UMBRAE-S7', our 'Finetuned' adaptation achieves comparable performance using only 30% of the data and often better when using more than 50%. Training only the tokenizer while keeping the pretrained backbone encoder frozen generally resulted in lower performance compared to fine-tuning the backbone together. This could be because the backbone encoder did not adequately incorporate the subject discrepancy in S7. Please consult the supplementary material for extra results and discussions on other subjects.

Table 5: Ablation Study. 'MLP' refers to the MLP-based brain encoder [37], while 'Enc-S' and 'Enc-U' represent our transformer-based encoders for single and multiple subjects. 'Dim.' means the output dimension of the brain encoder. The output needs to be passed to the adapter for further processing if it is 1024; otherwise, it is directly inputted into the LLM. The adapter has three training settings: 'Pretrained' means freezing the weights, 'Finetuned' means finetuning based on pretrained weights, and 'Joint' means training with the encoder from scratch. 'Loss Type' refers to loss functions (MSE, NCE, or both) applied to the outputs from the encoder (E.) or the adapter (A.).

Different Ablation Configurations				Captioning					Grounding			
									All		Salient	
Arch.	Dim.	Adapter	Loss Type	BLEU1	CIDEr	SPICE	CLIP-S	RefCLIP-S	acc@0.5	IoU	acc@0.5	IoU
MLP	1024	Pretrained	MSE (E.)	55.04	46.24	10.80	64.75	70.59	13.44	17.54	20.55	24.68
MLP	1024	Finetuned	MSE (A.)	54.02	43.24	10.35	64.09	70.02	13.56	17.91	20.92	25.54
Enc-S	1024	Pretrained	MSE (E.)	57.63	51.93	11.83	66.44	72.12	13.72	17.56	21.52	25.14
Enc-S	4096	N/A	MSE (A.)	52.06	36.40	9.06	62.30	68.27	13.31	17.04	20.85	24.78
Enc-S	1024	Joint	MSE (A.)	55.02	43.53	10.48	64.00	70.01	13.72	17.57	21.44	25.15
Enc-S	1024	Joint	MSE (E.) NCE (A.)	27.09	3.16	1.27	52.69	59.08	8.72	11.40	13.78	16.26
Enc-S	1024	Joint	MSE (A.) NCE (A.)	51.69	34.09	8.71	62.27	68.05	13.68	18.07	21.07	25.45
Enc-U	1024	Pretrained	MSE (E.)	59.44	61.06	12.79	67.78	73.54	18.93	21.28	30.23	30.18

5 Ablation Study

5.1 Architectural Improvements

MLP-based Encoder *vs.* Our Encoder. The MLP-based brain encoder is adapted from [37] with slight adjustments to match the desired output dimension. This deep MLP backbone amounts to 1,003.64 million parameters per subject. In comparison, our model needs only 112.63 million parameters for a single subject and 146.24 million for all four subjects. This translates to an 88.78% reduction in parameters for a single subject and a 96.36% reduction for all four subjects, respectively. Our single-subject encoder (denoted as UMBRAE-Sx) surpasses the MLP-based architecture [37] in captioning (Tab. 1), grounding (Tab. 2), and retrieval (Tab. 3) tasks by significant margins. The universal encoder (denoted as UMBRAE) achieves even greater performance improvements.

Single *vs.* Cross-Subject Design. The universal encoder differs from the single-subject encoder solely in the addition of subject-specific tokenizers (Sec. 3.1), and its training only varies in the batch sampling strategy (Sec. 3.2), enabling the training of diverse subjects within one model. Additionally, the resources required are basically the same as those for training a single-subject model, eliminating the necessity of extra training time or computational resources. This cross-subject design benefits from user diversity, achieving superior performance compared to focusing on a single subject. Results in Tabs. 1 to 3 show that the cross-subject design surpasses its single-subject counterpart across almost metrics.

5.2 Training Strategies

Current vision-language models typically comprise three main components: an image encoder, an adapter, and a large language model. Within this framework, there are several potential ways for multimodal-brain alignment. For example,

one could train the model to map brain responses to a pretrained semantic space through contrastive learning, which has been a common practice in previous methods. Alternatively, one could opt to fine-tune the adapter or train it jointly with the brain encoder, applying separate losses to each component. In this section, we delve into the motivation and rationale behind aligning with image features from the image encoder using a reconstruction loss. Further, in the supplementary we ablate our sampling strategy.

Pretrained *vs.* Finetuned Adapter. The adapter serves as the bridge connecting multimodal encoders [14,35] with the output space of MLLMs [7,27,44]. As shown in Tab. 5, either finetuning the adapter or training it jointly with the brain encoder results in decreased performance, likely due to catastrophic forgetting that occurs when updating well-trained parameters that have been learned from a significantly larger volume of data.

Image Feature *vs.* LLM Embedding. In addition to aligning with image features, we also conduct experiments on learning to align brain responses with the pretrained LLM used. Specifically, we explore three variants: (a) training the brain encoder and finetuning the adapter with different losses, (b) training the brain encoder and the adapter jointly, and (c) adjusting the output dimension of the brain encoder to align directly with the language embedding. As illustrated in Tab. 5, all attempts yield less desirable results compared to simply aligning with image features. Finetuning the entire model with new and sufficient data may indeed achieve better results, but acquiring such data can be challenging. However, experiments suggest that aligning with image features yields the best results when only image-brain pairs are available for training.

Reconstruction *vs.* Contrastive Learning. We further explore the effects of different loss functions on the aforementioned three variants, specifically a pixel-wise reconstruction loss (MSE) and a contrastive loss with MixCo augmentation (NCE) [30]. We also test applying separate losses to outcomes from the encoder and the adapter when trained jointly. The findings show that applying an MSE loss to the image feature while keeping the adapter unchanged leads to the most favorable performance in both concept recognition and object localization. Conversely, employing contrastive learning significantly diminishes performance.

6 Conclusion

In this work, we propose a method that decodes multimodal explanations from brain signals. Specifically, we introduce a universal brain encoder for multimodal brain alignment, which enables the recovery of conceptual and spatial details using multimodal large language models. To overcome unique brain patterns among different individuals, we introduce a novel cross-subject training strategy. This enables brain signals from multiple subjects to be trained within the same model and allows weakly-supervised subject adaptation, facilitating the training of a model for a new subject in a data-efficient manner. For evaluation, we construct BrainHub, a brain understanding benchmark, to facilitate future research.

Acknowledgements. This work was supported by the Engineering and Physical Sciences Research Council [grant EP/W523835/1] and the UKRI Future Leaders Fellowship [grant G104084].

References

1. Admoni, H., Scassellati, B.: Social eye gaze in human-robot interaction: a review. Journal of Human-Robot Interaction **6**(1), 25–63 (2017)
2. Allen, E.J., St-Yves, G., Wu, Y., Breedlove, J.L., Prince, J.S., Dowdle, L.T., Nau, M., Caron, B., Pestilli, F., Charest, I., et al.: A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. Nature neuroscience **25**(1), 116–126 (2022)
3. Anderson, P., Fernando, B., Johnson, M., Gould, S.: Spice: Semantic propositional image caption evaluation. In: ECCV. pp. 382–398. Springer (2016)
4. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: ACL Workshop. pp. 65–72 (2005)
5. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. NeurIPS **33**, 9912–9924 (2020)
6. Chaudhary, U., Vlachos, I., Zimmermann, J.B., Espinosa, A., Tonin, A., Jaramillo-Gonzalez, A., Khalili-Ardali, M., Topka, H., Lehmberg, J., Friehs, G.M., et al.: Spelling interface using intracortical signals in a completely locked-in patient enabled via auditory neurofeedback training. Nature communications **13**(1), 1236 (2022)
7. Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., Zhao, R.: Shikra: Unleashing multimodal llm's referential dialogue magic. arXiv preprint arXiv:2306.15195 (2023)
8. Chen, Z., Qing, J., Zhou, J.H.: Cinematic mindscapes: High-quality video reconstruction from brain activity. Advances in Neural Information Processing Systems **36** (2024)
9. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., et al.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. 2023. URL https://lmsys. org/blog/2023-03-30-vicuna **1**(2), 3 (2023)
10. Cui, Y., Zhang, Q., Knox, B., Allievi, A., Stone, P., Niekum, S.: The empathic framework for task learning from implicit human feedback. In: CoRL. pp. 604–626. PMLR (2021)
11. Desimone, R., Albright, T.D., Gross, C.G., Bruce, C.: Stimulus-selective properties of inferior temporal neurons in the macaque. Journal of Neuroscience **4**(8), 2051–2062 (1984)
12. Ferrante, M., Ozcelik, F., Boccato, T., VanRullen, R., Toschi, N.: Brain captioning: Decoding human brain activity into images and text. arXiv preprint arXiv:2305.11560 (2023)
13. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: CVPR (2023)
14. Han, J., Gong, K., Zhang, Y., Wang, J., Zhang, K., Lin, D., Qiao, Y., Gao, P., Yue, X.: Onellm: One framework to align all modalities with language. In: CVPR (2024)
15. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. In: EMNLP (2021)
16. Hussein, A., Gaber, M.M., Elyan, E., Jayne, C.: Imitation learning: A survey of learning methods. CSUR **50**(2), 1–35 (2017)
17. Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., Carreira, J.: Perceiver: General perception with iterative attention. In: ICML. pp. 4651–4664. PMLR (2021)
18. Kanwisher, N., McDermott, J., Chun, M.M.: The fusiform face area: a module in human extrastriate cortex specialized for face perception. Journal of neuroscience **17**(11), 4302–4311 (1997)

19. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of StyleGAN. In: CVPR. pp. 8107–8116 (2020)
20. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Communications of the ACM **60**(6), 84–90 (2017)
21. Laureys, S., Pellas, F., Van Eeckhout, P., Ghorbel, S., Schnakers, C., Perrin, F., Berre, J., Faymonville, M.E., Pantke, K.H., Damas, F., et al.: The locked-in syndrome: what is it like to be conscious but paralyzed and voiceless? Progress in brain research **150**, 495–611 (2005)
22. Lee, S., Zhang, R., Hwang, M., Hiranaka, A., Wang, C., Ai, W., Tan, J.J.R., Gupta, S., Hao, Y., Levine, G., et al.: Noir: Neural signal operated intelligent robots for everyday activities. In: CoRL (2023)
23. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: CVPR (2023)
24. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
25. Lin, S., Sprague, T., Singh, A.K.: Mind Reader: Reconstructing complex images from brain activities. NeurIPS **35**, 29624–29636 (2022)
26. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV. pp. 740–755 (2014)
27. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. NeurIPS **36** (2023)
28. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
29. Mai, W., Zhang, Z.: Unibrain: Unify image reconstruction and captioning all in one diffusion model from human brain activity. arXiv preprint arXiv:2308.07428 (2023)
30. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
31. Ozcelik, F., VanRullen, R.: Brain-Diffuser: Natural scene reconstruction from fMRI signals using generative latent diffusion. Scientific Reports **13**(1), 15666 (2023)
32. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: ACL. pp. 311–318 (2002)
33. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. In: ICLR (2024)
34. Puce, A., Allison, T., Asgari, M., Gore, J.C., McCarthy, G.: 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)
35. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PMLR (2021)
36. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
37. Scotti, P.S., Banerjee, A., Goode, J., Shabalin, S., Nguyen, A., Cohen, E., Dempster, A.J., Verlinde, N., Yundler, E., Weisberg, D., et al.: Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. In: NeurIPS (2023)
38. Smith, L.N., Topin, N.: Super-convergence: Very fast training of neural networks using large learning rates. In: Artificial intelligence and machine learning for multi-domain operations applications. vol. 11006, pp. 369–386. SPIE (2019)
39. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: CVPR. pp. 2818–2826 (2016)
40. Takagi, Y., Nishimoto, S.: High-resolution image reconstruction with latent diffusion models from human brain activity. In: CVPR. pp. 14453–14463 (2023)

41. Takagi, Y., Nishimoto, S.: Improving visual image reconstruction from human brain activity using latent diffusion models via multiple decoded inputs. arXiv preprint arXiv:2306.11536 (2023)
42. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: ICML. pp. 6105–6114. PMLR (2019)
43. Tang, J., LeBel, A., Jain, S., Huth, A.G.: Semantic reconstruction of continuous language from non-invasive brain recordings. Nature Neuroscience pp. 1–9 (2023)
44. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
45. Uddin, L.Q.: Salience processing and insular cortical function and dysfunction. Nature reviews neuroscience **16**(1), 55–61 (2015)
46. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NeurIPS **30** (2017)
47. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: CVPR. pp. 4566–4575 (2015)
48. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. TIP **13**(4), 600–612 (2004)
49. Xia, W., de Charette, R., Öztireli, C., Xue, J.H.: Dream: Visual decoding from reversing human visual system. In: WACV. pp. 8226–8235 (2024)
50. Xia, W., Yang, Y., Xue, J.H., Feng, W.: Controllable continuous gaze redirection. In: ACM MM. pp. 1782–1790 (2020)
51. Xu, X., Wang, Z., Zhang, E., Wang, K., Shi, H.: Versatile diffusion: Text, images and variations all in one diffusion model. In: ICCV. pp. 7754–7765 (2023)
52. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: ECCV. pp. 69–85. Springer (2016)