UMBRAE: Unified Multimodal Brain Decoding — Supplementary Material —

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

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

In this supplementary document, we provide further insights, experiments and analyses. First, in Sec. 1 we describe the tasks and their associated prompts used in UMBRAE and in Sec. 2 describe the BrainHub benchmark construction. We then report in Sec. 3 additional experiments to showcase the superiority of UMBRAE on all tasks, and extend our ablation and analyses in Sec. 4. Finally, in Sec. 5, we discuss the method limitations and potential negative impacts.

1 UMBRAE: Tasks and Example Prompts

Our method inherits multimodal understanding capabilities of MLLMs, enabling the switch between different tasks through the use of various task prompts. Taking Shikra [4] as an example, we excerpts the task prompts used during their training process in Tab. 1. Three prompts for each task are randomly selected to provide readers an intuitive understanding. During inference, users are not constrained to these predefined formats and are free to frame their queries in natural language, allowing for a broad spectrum of diverse and compelling task formats.

It should be noted that our method is compatible with nearly all tasks featured in Shikra [4] and LLaVA [13], with the exception of tasks that necessitate initially locating an object or scene as inputs, such as grounding captioning and referring expression generation. Importantly, note that although the term 'image' is used in the prompts we do *not* utilize images as input in our brain understanding tasks. *The reference images are only used for visualization purposes.* In practice, the prompt embedding will be concatenated with the brain features predicted from our brain encoder, utilizing brain signals as inputs. We report detailed results for each task of Tab. 1 later in Sec. 3.

2 Details on BrainHub

Our multimodal brain understanding benchmark, coined BrainHub, extends the popular NSD [1] using COCO [12] annotations. Here, we first describe NSD (Sec. 2.1) and then elaborate the construction of BrainHub (Sec. 2.2).

2.1 Natural Scenes Dataset

Natural Scenes Dataset [1] (NSD) stands as the largest publicly accessible fMRI dataset. It encompasses brain activity recordings from 8 subjects (participants)

Table 1: Supported Task Prompts. The tags *<image>*, *<question>*, and *<expr>* are placeholders, representing input images, questions in QA tasks, and expressions in the REC task. During inference, users are free to create diverse task formats according to actual needs. Q, A, C, and C^{Box} denote the **Q**uestion, **A**nswer, **C**hain of thoughts (CoT), and **C**oT with **Box**. CoT is delivering an answer along with the reasoning process. 'Box' denotes coordinates of bounding boxes.

Task	Three randomly chosen examples from hundreds.
Captioning	Describe this image <image> as simply as possible. What is the content of the image <image>? Please answer in short sentences. Summarize the content of the photo <image>.
REC	In the given <image>, could you find and tell me the coordinates of <expr>? I need the coordinates of <expr> in <image>, can you please assist me with that? Locate <expr> in <image> and provide its coordinates, please.
Spotting Cap.	Can you provide a description of the image <image> and include the coordinates [x0,y0,x1,y1] for each mentioned object? Please explain what's happening in the photo <image> and give coordinates [xmin,ymin,xmax,ymax] for the items you reference. How would you describe the contents of the image <image>? Please provide the positions of mentioned objects in square brackets.
Q → A	I want to know the answer to '<question>' Refer to the image <image> and give a clear response. Answer this question directly after referring to the image <image>: <question> Examine the image <image> and provide a brief answer for '<question>'
Q → CA	Having a look at image <image>, can you tell me the answer to my question '<question>' and the logic leading to it? Please answer the following question '<question>' based on the image <image>, and describe your thought process Upon analyzing the image <image>, please find the answer to my question '<question>' and provide a detailed explanation.
Q→ C^{Box}A	<question> Please offer your reasoning process, and provide bounding boxes of mentioned objects within square brackets. Here is the picture <image> Please explain your reasoning and provide bounding boxes, denoted by square brackets, for the objects mentioned in the picture <image>. <question> Consider <image>, and then provide a well-reasoned answer to '<question>' Don't forget to mark relevant object locations using [x0,y0,x1,y1].

who viewed images passively for up to 40 hours inside an MRI machine. Each image was displayed for three seconds and repeated three times across 30-40 scanning sessions, yielding 22,000–30,000 fMRI response trials per subject.

Following recent studies, we utilize the four subjects who finished all scanning sessions, that is: S1, S2, S5, and S7. As in [20, 22, 25], we utilize preprocessed voxels corresponding to the 'nsdgeneral' brain region. The latter region, described by the NSD authors, comprises the subset of voxels in the posterior cortex most responsive to the presented visual stimuli. In the training set for each subject, there are 8,859 images and 24,980 fMRI trials (with each image tested up to three times). We compute the average response as per previous studies [20]. The test set comprises an additional 982 images and 2,770 fMRI trials common across four individuals. Importantly, all images used during the fMRI recordings are from the COCO [12] dataset, which we conveniently use to retrieve the original COCO labels to construct our BrainHub benchmark.

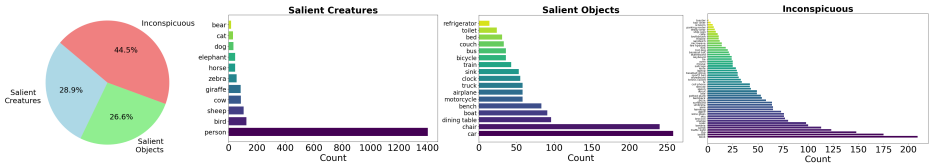
2.2 Benchmark Construction

BrainHub extends NSD [1] for the four subjects who finished all scanning sessions (S1, S2, S5, and S7). Tab. 2 outlines the characteristics of the benchmark. Each image viewed by the subject contains several captions and may include multiple bounding boxes (instances) of each class. According to their salience in images [24], we group the 80 classes of COCO into 4 salience categories: Salient, Salient Creatures, Salient Objects, and Inconspicuous. Note that 'Salient' is the combination of Salient Creatures and Objects. Fig. 1 shows the statistics and mapping of our categories, w.r.t. COCO classes. The inconspicuous (**I**) category accounts for the largest proportion, while the salient objects (**SO**) and creatures (**SC**) are roughly even. The classes within each category exhibit significant

imbalance; for instance, the number of instances for the 'person' class stands out in the salient category.

Table 2: BrainHub Details. The test set characteristics include the number of images, voxels, captions, bounding boxes, regions of interest (ROIs) in the fMRI data, subject references, and their corresponding dimensions.

Images	Classes	Captions	Bounding boxes	ROIs	Subject	Dimension
982	80	4,913	5,829	V1, V2, V3, hV4, VO, PHC, MT, MST, LO, IPS	S1 S2 S5 S7	15,724 14,278 13,039 12,682



(a) Statistics.

Category	COCO classes (# of classes)
A	**S** + **I** (80)
S	**SC** + **SO** (28)
SC	person, bird, cat, dog, horse, sheep, cow, elephant, bear, zebra, giraffe (11)
SO	bicycle, car, motorcycle, airplane, bus, train, truck, boat, bench, chair, couch, bed, dining table, toilet, sink, refrigerator, clock (17)
I	traffic light, fire hydrant, stop sign, parking meter, backpack, umbrella, handbag, tie, suitcase, frisbee, skis, snowboard, sports ball, kite, baseball bat, baseball glove, skateboard, surfboard, tennis racket, bottle, wine glass, cup, fork, knife, spoon, bowl, banana, apple, sandwich, orange, broccoli, carrot, hot dog, pizza, donut, cake, potted plant, tv, laptop, mouse, remote, keyboard, cell phone, microwave, oven, toaster, book, vase, scissors, teddy bear, hair drier, toothbrush (52)

(b) Mapping between categories and COCO classes.

Fig. 1: BrainHub Statistics. We illustrate the statistics in (a) and the mapping relationships in (b) for the categories ontology used in BrainHub, w.r.t. to the original COCO classes. Please zoom in (a) for details.

3 Additional Experiments

3.1 Brain Captioning

More Results for S1. In addition to the quantitative evaluation of brain captioning on S1 reported in the main paper, we report a few qualitative comparison in Fig. 2. As shown, SDRecon [22] often produces incoherent and irrelevant

descriptions, consistent with the lowest fluency and relevance metrics in the quantitative evaluation results. OneLLM [7] provides complete and coherent responses, but typically diverges from the content in the reference visual stimuli, exhibiting the lowest CLIP similarity scores [8] (CLIP-S and RefCLIP-S). BrainCap [6] yields better results compared to the above two methods, with fluency not being an issue but encountering challenges in content similarity with the reference.

Both of our methods, UMBRAE and UMBRAE-S1, yield results with the most fluent and relevant descriptions. Here, 'UMBRAE-Sx' refers to our model trained with a single subject only and 'UMBRAE' is the unified brain decoding model with cross-subject training. When other methods fail to provide accurate descriptions, ours can accurately describe the scenes, such as parasailing, lake, mountain in row 1; building with clock tower in row 2; bird on a tree branch in row 5, skiing in the last row. Besides, baselines tend to provide approximate descriptions in contradiction to our method, such as for giraffe in row 3, sheep in row 6 (note UMBRAE even provides accurate counts), living room in the second last. Importantly, note that unlike the brain captioning SOTAs [6, 7, 14, 22], we do *not* use any ground truth captions during the training of the brain encoder.

Results for S2, S5, S7. Tab. 3 shows the brain captioning results on S2, S5, S7. We compare our single-subject model UMBRAE-Sx and cross-subject model UMBRAE with two state-of-the-art methods SDRecon [22] and BrainCap [6]. Results for other subjects in OneLLM [7] and UniBrain [14] are unavailable: OneLLM [7] solely trains with S1, and UniBrain [14] has not been open-sourced, with no reported results for other subjects. The results on S2, S5, and S7 show consistent performance with those presented for S1 in the main paper. Remarkably, in all metrics and subjects, our methods perform better than the baselines [6, 22]. The models using cross-subject training (UMBRAE) also generally perform better than those trained on a single subject (UMBRAE-Sx).

Subject Comparison. Fig. 3 shows brain captioning results on different subjects. The first lines are ground truth captions from COCO [12] for comparison. When an image comes with multiple ground truth captions, we always select the one. The second lines display the results from Shikra [4] using images as input, which can be seen as an approximate upper bound of performance for our method. Our method shares the same adapter and the finetuned LLM with Shikra in most experiments, but uses a different encoder (brain encoder instead of image encoder) and input modality (brain response instead of image). The following lines show the results for different subjects (S1, S2, S5, and S7). Results for these subjects are all from a single UMBRAE model with subject-specific training using only the brain responses as input. Our method achieves comparable quality and relevance even when compared to ground truth, or the image captioning method using images as input.

3.2 Brain Grounding

Results of acc@0.3 and acc@0.7 on different subject (S1, S2, S5, and S7) and different categories (A, S, SC, SO, and I) are reported in Tab. 4. Results of

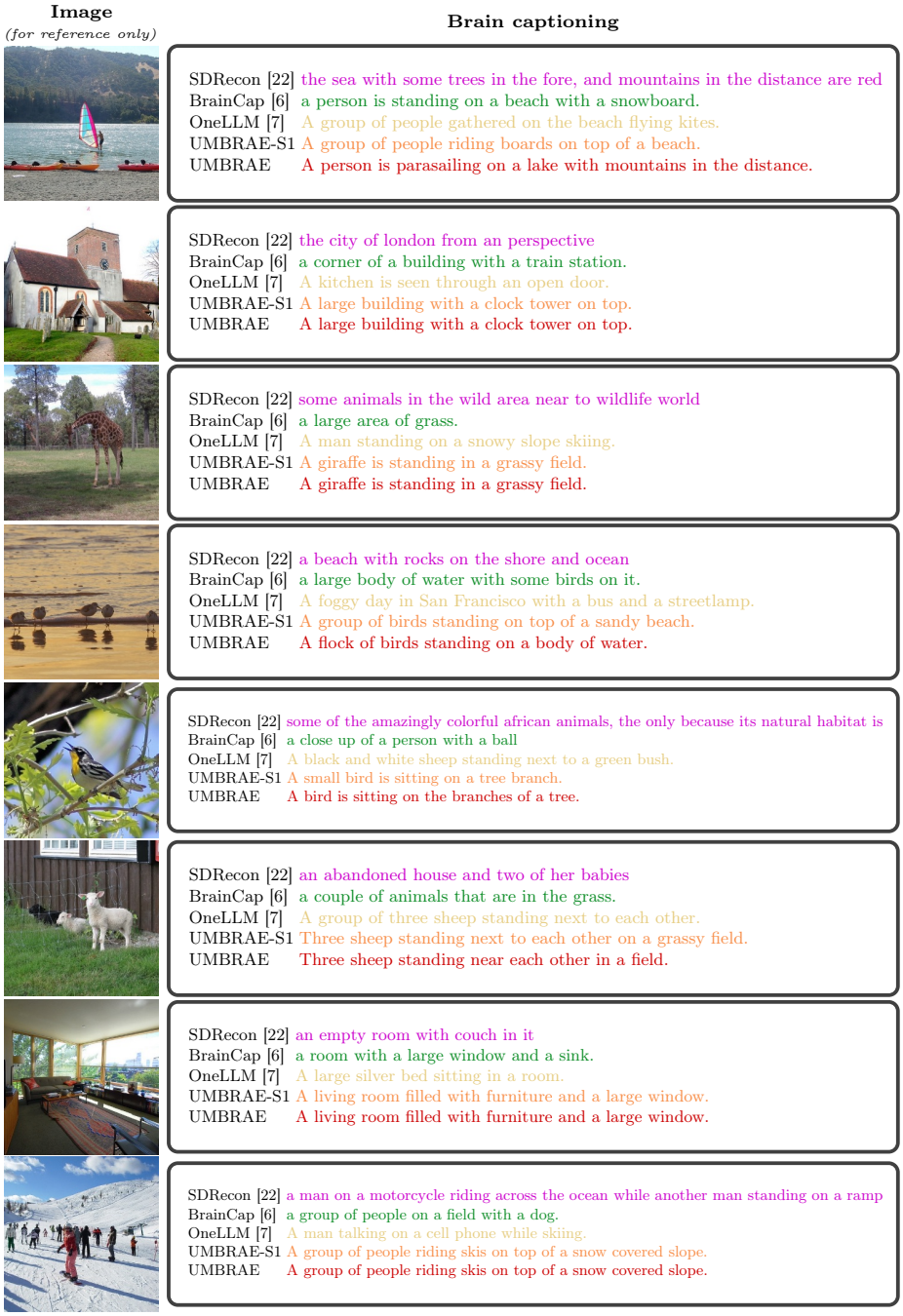


Fig. 2: Brain Captioning Comparison on S1. Baselines for S1 include SDRecon [22], BrainCap [6], and OneLLM [7]. 'UMBRAE-S1' refers to our model trained only with subject S1, while 'UMBRAE' denotes the model with cross-subject training.

Table 3: Brain Captioning. 'UMBRAE-Sx' refers to our model trained with a single subject only, while UMBRAE' denotes the model with cross-subject training. The results of S1 have been presented in the main paper and are listed here for completeness. The colors represent the best, second-best, and third-best performance. Note that both our models are consistently surpassing the baselines and, furthermore, our cross-subject model (UMBRAE) is almost always better than its single-subject counterpart.

Method	Eval	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE	CIDEr	SPICE	CLIP-S	RefCLIP-S
SDRecon [21]	S1	36.21	17.11	7.72	3.43	10.03	25.13	13.83	5.02	61.07	66.36
BrainCap [6]		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
SDRecon [22]	S2	34.71	15.87	6.72	3.02	9.60	24.22	13.38	4.58	59.52	65.30
BrainCap [6]		53.80	34.16	20.86	13.03	15.90	39.96	35.60	8.47	62.48	68.19
UMBRAE-S2		57.18	37.76	25.06	17.18	18.11	41.85	50.62	11.50	64.87	71.06
UMBRAE		59.37	40.47	27.14	18.41	19.17	43.86	55.93	12.08	66.46	72.36
SDRecon [22]	S5	34.96	16.39	7.36	3.49	9.93	24.77	13.85	5.19	60.83	66.30
BrainCap [6]		55.28	35.71	22.62	14.62	16.45	40.87	41.05	9.24	63.89	69.64
UMBRAE-S5		58.99	39.88	27.03	18.73	19.04	43.30	57.09	12.70	66.48	72.69
UMBRAE		60.36	41.27	27.92	19.03	20.04	44.81	61.32	13.19	68.39	74.11
SDRecon [22]	S7	34.99	16.10	7.06	3.26	9.54	24.33	13.01	4.74	58.68	64.59
BrainCap [6]		54.25	34.47	21.67	14.00	15.94	40.02	37.49	8.57	62.52	68.48
UMBRAE-S7		55.71	36.24	23.62	15.75	17.51	40.64	47.07	11.26	63.66	70.09
UMBRAE		57.20	38.30	25.49	17.13	18.29	42.16	52.73	11.63	65.90	71.83

acc@0.5 are reported in the main paper. The IoU values remain consistent with those presented in the main paper.

Generally, the cross-subject model outperforms the single-subject models. These results are obtained using the grounding prompt '`Locate <expr> in <image> and provide its coordinates, please.`', where `<expr>` is the expression. In practice, to obrain the evaluation results shown in Tab. 4, we assume there are several 'known' concepts (`<expr>`) in the given image and formulate the task as detecting queried objects and returning their coordinates. This is known as the REC task in MLLMs. The example supported task prompts are shown in the second row of Tab. 1.

Brain Grounding without Priors. Interestingly, our method can also provide descriptions and coordinates for brain signals without 'prior knowledge' of their contents. This task is referred to as 'Spotting Captioning' in [4], but it is a grounding-related task, as the primary goal is to describe the image (brain responses in our case) and identify the mentioned objects or regions using points or boxes. In addition to the prompts shown in Tab. 1, we can also utilize a wide range of instructions such as '`Please interpret this image and give coordinates [x1,y1,x2,y2] for each object you mention`' and '`Provide a detailed description of the image using around 100-500 words, including the objects, attributes, and spatial locations depicted in the picture`'. The qualitative results of brain grounding, using various task prompts and across different subjects, are provided in Fig. 4 and Fig. 5, respectively. The bounding boxes are outputted as text responses, and we visualize the outputs in the reference images. The tags `<expr>` in the REC prompt are depicted in the corresponding reference images with color. The concepts and coordinates in the

Image *(for reference only)*	Brain captioning	
	COCO [12]	A bathroom with a vanity mirror sitting above a toilet next to a bathtub.
	Shikra-w/img [4]	A bathroom with a toilet, sink and a television.
	S1	A bathroom with a toilet, sink and mirror.
	S2	A bathroom with a sink, mirror and toilet.
	S5	A kitchen with a stove, sink, and cabinets
	S7	A bathroom with a toilet, sink and bathtub.
	COCO [12]	A picture of a cat and some luggage.
	Shikra-w/img [4]	A cat sitting on a suitcase with clothes on a table.
	S1	A cat is sitting on top of a closed suitcase.
	S2	A cat is laying down on a soft surface.
	S5	A cat is laying on top of a bed.
	S7	A cat laying on top of a bed in a room.
	COCO [12]	A large field of grass with sheep grazing on the land.
	Shikra-w/img [4]	A herd of sheep graze in a lush green field.
	S1	A large mountain range filled with lots of trees.
	S2	The image shows a great wilderness of mountains.
	S5	A large mountain range is shown with a sky in the background.
	S7	A large field with a mountain range in the background.
	COCO [12]	A man riding a snowboard down a hill.
	Shikra-w/img [4]	A skier is going down a snowy hill.
	S1	A person in a ski outfit skiing down a slope.
	S2	A man riding a surfboard on top of a wave.
	S5	A person on skis is skiing on a snowy slope.
	S7	A person riding a snowboard on top of a snow covered slope.
	COCO [12]	Double decker bus on the street next to buildings.
	Shikra-w/img [4]	A double decker bus is parked outside a building.
	S1	A transit bus riding down a street with buildings around.
	S2	A passenger bus that is driving down the street.
	S5	A large bus is traveling down the street.
	S7	A bus driving down the street near another bus.
	COCO [12]	A person holding a tennis racket in their hand.
	Shikra-w/img [4]	A young man in an orange shirt playing tennis.
	S1	A woman holding a tennis racquet on top of a tennis court.
	S2	A woman holding a tennis racket on a tennis court.
	S5	A woman standing on a tennis court holding a racket.
	S7	A man holding a tennis racquet on a tennis court.

Fig. 3: Brain Captioning Results on Different Subjects. 'COCO' is the ground truth caption in the COCO dataset [12]. We excerpt the first caption if there are multiple captions for the same image. Shikra-w/img [4] is the result using the ground truth images as input. Results for all four subjects (S1, S2, S5, and S7) are from our cross-subject UMBRAE model.

Table 4: Brain Grounding Results of acc@0.3 and acc@0.7 on Different Subjects. The accuracy with threshold m is abbreviated as acc@m. The IoU values remain consistent with those presented in the main paper, corresponding to each subject and category. 'UMBRAE-Sx' refers to our model trained with a single subject only, while 'UMBRAE' denotes the model with cross-subject training. Results of acc@0.5 are reported in the main paper. The best results per subject is in color.

Method	Eval	All (A)		Salient (S)		Salient Creatures (SC)		Salient Objects (SO)		Inconspicuous (I)	
		acc@0.3	acc@0.7	acc@0.3	acc@0.7	acc@0.3	acc@0.7	acc@0.3	acc@0.7	acc@0.3	acc@0.7
UMBRAE-S1	S1	24.22	5.75	36.26	9.08	43.71	10.00	28.15	8.09	9.20	1.58
UMBRAE		30.47	8.47	44.45	13.55	55.86	16.14	32.04	10.73	13.01	2.14
UMBRAE-S2	S2	26.00	6.53	39.09	10.35	47.43	11.43	30.02	9.18	9.67	1.77
UMBRAE		29.60	7.94	42.96	12.58	55.14	16.00	29.70	8.86	12.92	2.14
UMBRAE-S5	S5	26.04	5.99	39.09	9.23	45.57	9.14	32.04	9.33	9.76	1.95
UMBRAE		30.05	7.28	44.75	11.47	56.14	14.29	32.35	8.40	11.71	2.04
UMBRAE-S7	S7	25.47	5.21	37.97	8.12	46.43	7.86	28.77	8.40	9.85	1.58
UMBRAE		28.32	7.03	42.07	10.80	53.14	12.86	30.02	8.55	11.15	2.32

responses of the Spotting Captioning task are depicted using the same color as the bounding box color in the visualizations.

3.3 Brain Retrieval

Fig. 6 shows the forward and backward retrieval results using MindEye [20] and our UMBRAE. The images displayed in the top row are the reference image and the top 5 retrieval images obtained from the *forward* retrieval [11]. This process is to find the correct paired CLIP image embedding given a brain embedding. Similarly, the bottom row are the reference image and the top 5 retrieval images obtained from the *backward* retrieval process, which aims to locate the correct brain embedding given an image embedding.

3.4 Visual Decoding

Despite that our method is not specifically designed for the task of visual decoding (fMRI-to-image reconstruction), our predicted textual and visual cues can be used for the final image reconstruction by using a variety of pretrained image generation models. To be specific, we use three text-to-image models SD [19], SD-XL [17] and Kandinsky [2], a layout-to-image model GlIGEN [10], and a multiple-condition model Versatile Diffusion (VD) [26]. The quantitative evaluation results are in Tab. 5. Besides the common visual decoding metrics [15], we further evaluate the image quality using FID [9] and Clean-FID (CFID) [16], as well as the image similarity using LPIPS [27]. The qualitative comparison is illustrated in Fig. 8.

Given that our method is VD-based [26], we further analyze the effects of its inputs, as shown in the last three rows in Tab. 5. For the dual context mode (generating an image conditioned on both text and image) of VD [26], three parts can be used as inputs for the visual decoding task: the input text (t), the input image (i), and a latent code z that can either be encoded from a given image or randomly sampled. We test the upper limits of performance for different inputs.

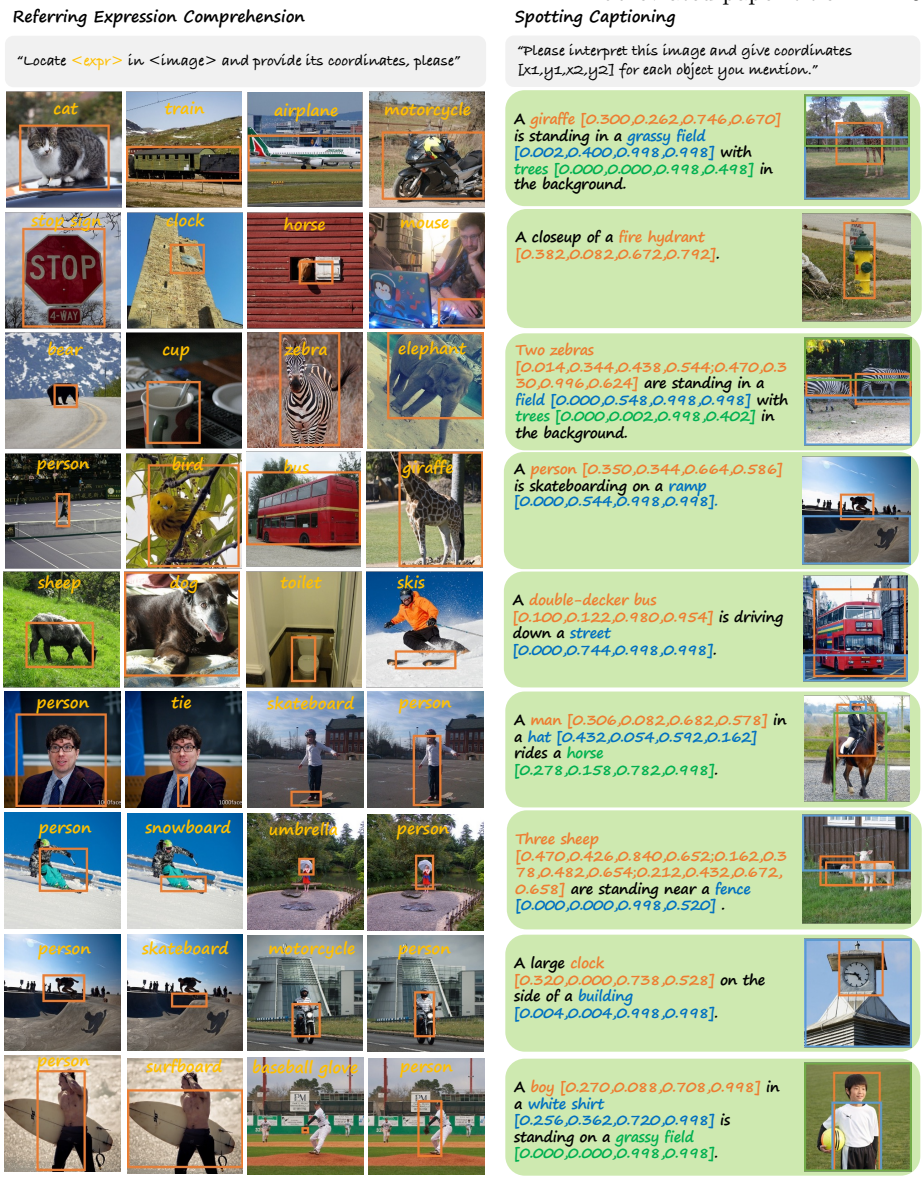


Fig. 4: Brain Grounding Results using Different Task Prompts. The shown results are from our UMBRAE using brain responses as input. Reference images are visual stimuli for input brain responses and are just used here for visualization. The tags `<expr>` in the REC prompt are depicted in the corresponding reference images. The bounding boxes are outputted as text responses, and we visualize the outputs in the reference images. The tags `<expr>` in the REC prompt are displayed in the reference images with color. The concepts and their coordinates in the Spotting Captioning responses are color-coded to match the bounding box color in the visualizations.

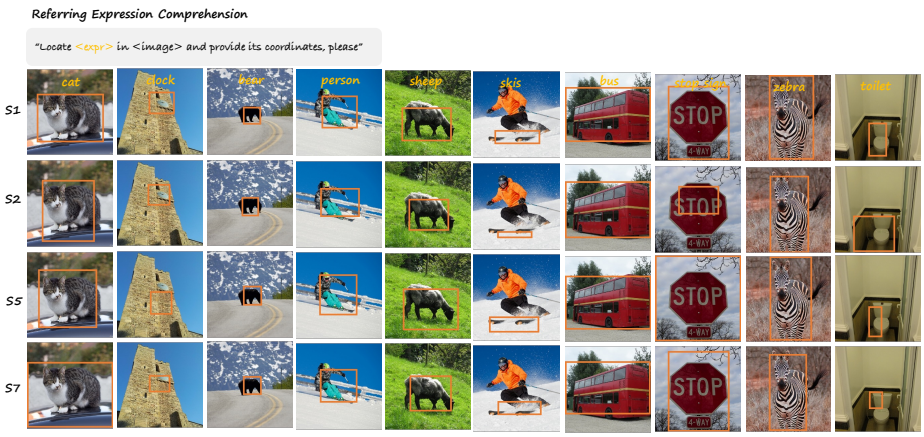
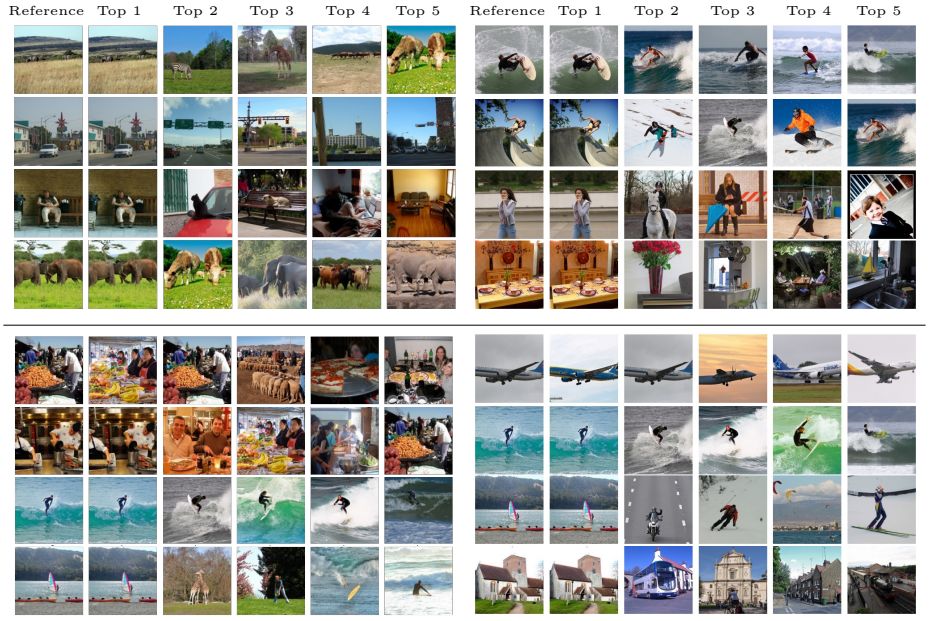


Fig. 5: Brain Grounding Result on Different Subjects. All results are from our UMBRAE using brain responses as input. Reference images here are for visualization.



(a) MindEye [20] **(b)** UMBRAE

Fig. 6: Brain Retrieval Results. We show the forward and backward retrieval results of MindEye [20] and our UMBRAE. The forward retrieval [11] (**top**) is to find the correct image embedding given a brain embedding. Conversely, the backward retrieval (**bottom**) aims to locate the correct brain embedding given an image embedding.

Specifically, VD-1 is using ground truth image and text, along with the latent code decoded from the ground truth image. VD-2 is similar but uses randomly sampled latent codes. VD-3 uses ground truth image and text, as well as predicted low-level images from Brain-Diffuser [15]. As shown in Tab. 5, latent codes z in the latent diffusion-based image generation models play a crucial role in the final visual decoding performance. Simply using predicted text (SD, SDXL, and Kandinsky3) or adding predicted image embedding (VD-2) is not sufficient for reliable visual decoding results. The predicted bounding boxes (GLIGEN) are intended to function as low-level constraints for the final reconstruction. However, they do not perform well on low-level metrics due to their unreliable nature.

Fig. 9 displays a qualitative comparison of individual subjects, while the quantitative evaluation of UMBRAE on all four subjects can be found in Tab. 6. Fig. 7 shows comparison on visual decoding between our method and the literature.

Fig. 7: Visual Decoding Comparison between UMBRAE and the literature on NSD. All reconstructed images are from S1.

Table 5: UMBRAE Visual Decoding for S1 with Various Image Generation Models. Despite that our method is not specifically designed for fMRI-to-image reconstruction (visual decoding), the predicted textual and visual outputs can serve as cues for the final image reconstruction using a variety of pretrained image generation models. These models include text-to-image SD [19], SD-XL [17], and Kandinsky [2], a layout-to-image GlIGEN [10], and a multiple-condition Versatile Diffusion (VD) [26]. We further analyze the effects of VD's inputs in the last three rows. The colors represent the best, second-best, and third-best performance.

Image Generation	Low-Level				High-Level				FID ↓	CFID ↓	LPIPS ↓
	PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	EffNet-B ↓	SwAV ↓			
SD [19]	-	.292	71.7%	83.5%	85.6%	86.1%	.786	.535	95.62	89.04	0.79
SDXL [17]	.070	.336	73.6%	86.9%	87.2%	86.3%	.769	.475	85.67	82.41	0.76
Kandinsky3 [2]	.110	.328	75.9%	85.4%	85.8%	86.4%	.789	.514	89.98	88.01	0.78
GLIGEN [10]	.078	.255	78.6%	90.1%	86.5%	87.5%	.766	.473	73.70	67.21	0.77
VD [26]	.293	.345	95.8%	97.2%	92.6%	93.9%	.690	.391	67.47	63.81	0.73
VD-1 (GT-ti-z)	.641	.402	99.9996%	99.9985%	99.5%	99.98%	.390	.187	54.33	44.74	0.53
VD-2 (GT-ti-rand-z)	.098	.244	91.0%	98.9%	98.8%	99.93%	.504	.261	59.12	53.30	0.69
VD-3 (GT-ti-pred-z)	.327	.352	98.975%	99.705%	98.9%	99.90%	.497	.261	62.96	52.48	0.71

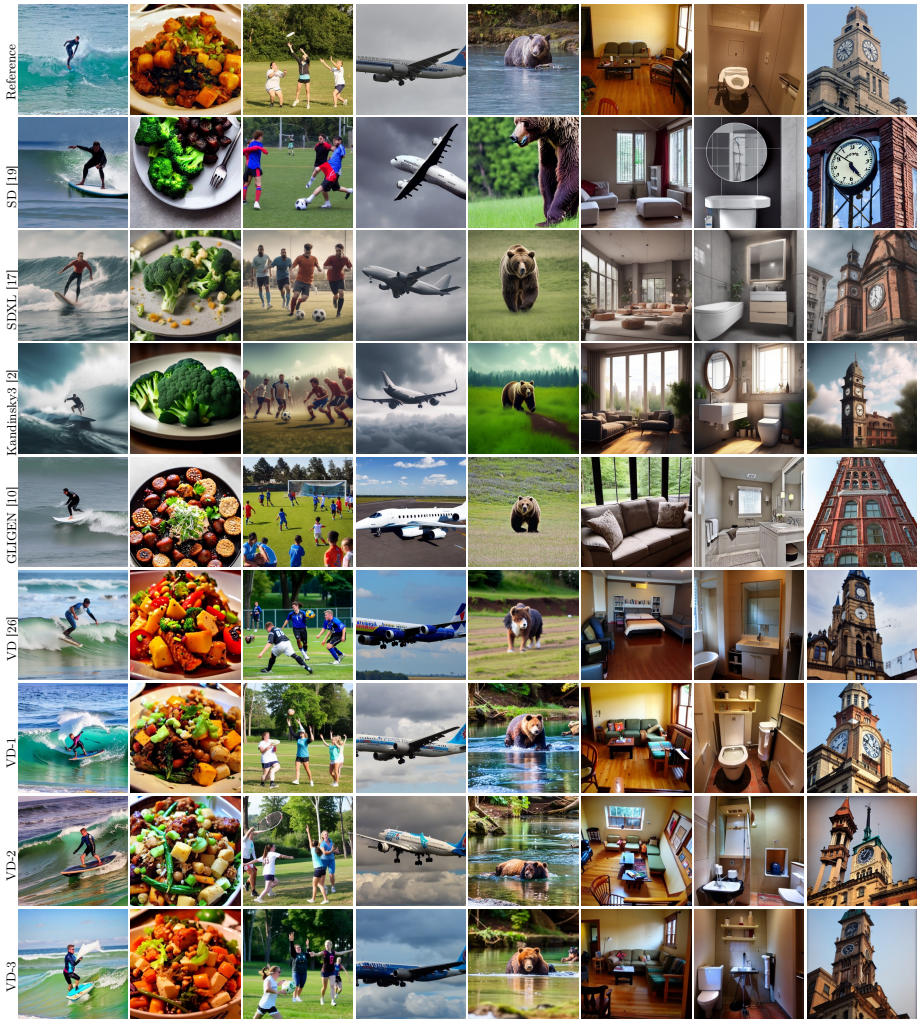


Fig. 8: UMBRAE Visual Decoding with Various Image Generation Models. We use textual and visual outputs predicted by our UMBRAE model as cues for the final image reconstruction, using a variety of pretrained image generation models. These models include text-to-image SD [19], SD-XL [17], and Kandinsky3 [2], a layout-to-image GlIGEN [10], and a multiple-condition VD [26]. Given that our method is based on VD [26], we further analyze the effects of its inputs. VD-1 is using ground truth image and text, along with the latent code decoded from the ground truth image. VD-2 is similar but uses randomly sampled latent codes. VD-3 uses ground truth image and text, as well as predicted low-level images from Brain-Diffuser [15].

Table 6: Subject-Specific Visual Decoding Evaluation. Quantitative evaluation of the UMBRAE reconstruction for the four subjects (S1, S2, S5, and S7) of NSD [1].

Subject	Low-Level				High-Level						
	PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	EffNet-B ↓	SwAV ↓	FID ↓	CFID ↓	LPIPS ↓
S1	.293	.345	95.8%	97.2%	92.6%	93.9%	.690	.391	67.47	63.81	0.73
S2	.283	.353	96.2%	97.3%	90.8%	93.0%	.705	.396	70.00	66.78	0.74
S5	.277	.337	95.8%	97.7%	93.7%	94.8%	.689	.380	67.70	64.52	0.74
S7	.279	.329	95.0%	96.9%	90.5%	92.1%	.713	.405	70.07	66.50	0.75

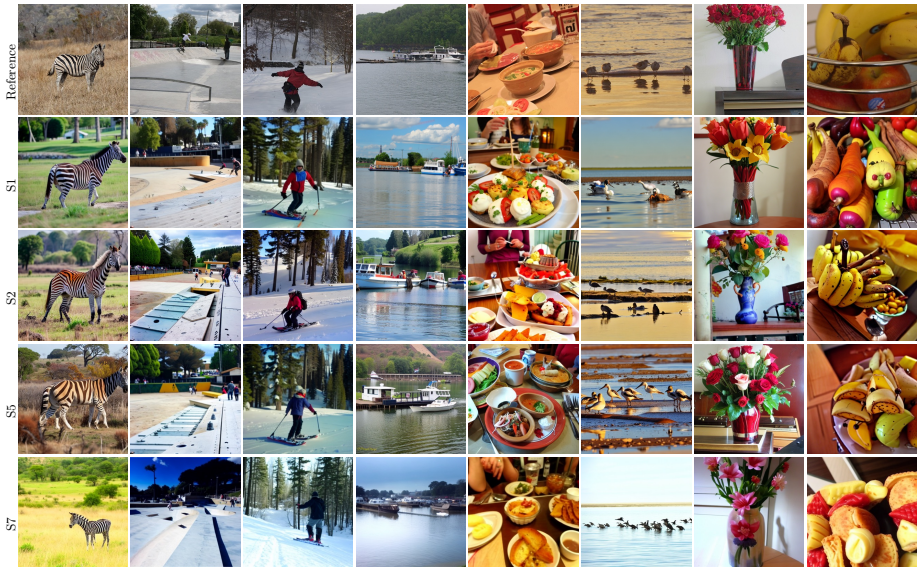


Fig. 9: Subject-Specific Visual Decoding Results. Qualitative comparison of UMBRAE reconstruction for the four subjects (S1, S2, S5, and S7) of NSD [1].

4 Additional Analyses

Here, we provide a couple of additional studies to further evaluate the per-subject performance (Sec. 4.1), the weakly-supervised adaptation (Sec. 4.2). We analyze different sampling strategies in cross-subject training (Sec. 4.3) and present a joint grounding-decoding evaluation (Sec. 4.4). We also build on MLLMs capacities to explore other tasks (Sec. 4.5) and finally, demonstrate the model-agnostic characteristics of our method (Sec. 4.6).

4.1 Subject-Specific Analysis

In the main paper, we primarily present the results from two types of our models: the cross-subject training model UMBRAE and the single-subject training model UMBRAE-Sx. The former is trained using all training data from subjects {S1, S2, S5, S7}, whereas UMBRAE-Sx trains only on data from a specific subject.

Table 7: UMBRAE Subject Analysis. We report the per-subject evaluation of UMBRAE when varying the set of training subjects (col 'Train'). For example, 'S1' denotes our method is trained only on data from S1, *i.e.* this corresponds to UMBRAE-S1. Similarly, 'S1,S2,S5,S7' means training using data from all subjects, and corresponding to the model UMBRAE. 'S1,S2,S5' means the model is trained with samples from {S1, S2, S5}. 'Eval' means evaluation on a certain subject. We note that better performance are almost always achieved when training on more than one subject. The colors represent the best, second-best, and third-best performance.

UMBRAE setting		Captioning						Grounding			
Train	Eval	BLEU1	ROUGE	CIDEr	SPICE	CLIP-S	RefCLIP-S	acc@0.5 (A)	IoU (A)	acc@0.5 (S)	IoU (S)
S1		57.63	42.15	51.93	11.83	66.44	72.12	13.72	17.56	21.52	25.14
S1,S2,S5		59.75	43.53	57.36	12.79	66.39	72.63	15.63	19.30	23.60	27.15
S1,S2,S7	S1	58.31	42.67	55.03	12.20	65.72	71.95	14.76	18.76	22.93	26.58
S1,S5,S7		59.23	43.70	57.00	12.69	66.25	72.42	15.58	19.34	24.05	27.58
S1,S2,S5,S7		59.44	43.71	61.06	12.79	67.78	73.54	18.93	21.28	30.23	30.18
S2		57.18	41.85	50.62	11.50	64.87	71.06	15.21	18.68	23.60	26.59
S1,S2,S5		57.91	42.43	52.80	11.91	65.37	71.57	16.04	19.29	25.09	27.43
S1,S2,S7	S2	57.30	41.98	51.17	11.42	64.53	70.85	15.42	19.27	24.05	27.60
S2,S5,S7		57.69	42.29	51.77	11.72	65.08	71.27	15.25	18.95	23.08	26.70
S1,S2,S5,S7		59.37	43.86	55.93	12.08	66.46	72.36	18.27	20.77	28.22	29.19
S5		58.99	43.30	57.09	12.70	66.48	72.69	14.72	18.45	22.93	26.34
S1,S2,S5		59.63	43.32	60.00	13.25	67.10	73.16	15.01	18.90	23.60	26.97
S1,S5,S7	S5	60.02	43.50	59.67	13.31	67.24	73.39	14.84	18.68	22.86	26.34
S2,S5,S7		59.43	43.45	59.22	12.71	67.10	73.24	15.05	18.82	23.16	26.52
S1,S2,S5,S7		60.36	44.81	61.32	13.19	68.39	74.11	18.19	20.85	28.74	30.02
S7		55.71	40.64	47.07	11.26	63.66	70.09	13.60	17.83	21.07	25.19
S1,S2,S7		56.72	41.43	49.78	11.37	64.21	70.62	14.43	18.26	21.82	25.64
S1,S5,S7	S7	57.83	42.09	53.53	11.88	64.92	71.29	14.76	18.68	22.49	26.26
S2,S5,S7		56.29	41.64	51.23	11.52	64.46	70.84	14.43	18.22	22.11	25.68
S1,S2,S5,S7		57.20	42.16	52.73	11.63	65.90	71.83	16.74	19.63	25.69	27.90

Instead, here we explore training the cross-subject models with various combinations of different subjects (*e.g.*, training on {S1, S2, S5} and testing on S1). These pretrained models are then utilized in the weakly-supervised subject adaptation experiments. The corresponding results are detailed in Tab. 7. It is interesting to note that while training with more than one subject is always improving performance, the best performance are not always achieved when using all subjects available. An example is the S7 evaluation which performs better when UMBRAE is trained only on {S1, S5, S7} (*i.e.*, not using S2 data). We conjecture this could relate to some subjects having more similar brain activities patterns than others.

4.2 Weakly-Supervised Adaptation

As explained in the main paper, our method enables weakly-supervised subject adaptation and can train a model for a new subject in a data-efficient manner. In other words, UMBRAE can accommodate a new subject with only a portion of the total training data. This is crucial considering the challenges in obtaining brain modality data. In the main paper, we presented the results of pretraining on {S1, S2, S5} and adapting to S7 by finetuning with varying amounts {5%, 10%, 20%, 30%, 50%, 80%, 100%} of S7 training data.

In Fig. 10, we present the adaptation to other subjects, being: **Fig. 10a** training on {S2, S5, S7}, adaptation to S1; **Fig. 10b** training on {S1, S5, S7}, adaptation to S2; **Fig. 10c** training on {S1, S2, S7}, adaptation to S5. In the above mentioned figures, we report performance when finetuning both the tokenizer and the encoder, which was proven to provide the best performance in the main paper. As for adaptation to S7 (cf. main paper), compared to the single-subject model 'UMBRAE-Sx', our 'Finetuned' adaptation on S1, S2, and S5 achieves comparable performance using only 30% of the data.

4.3 Sampling Strategies in Cross-Subject Training

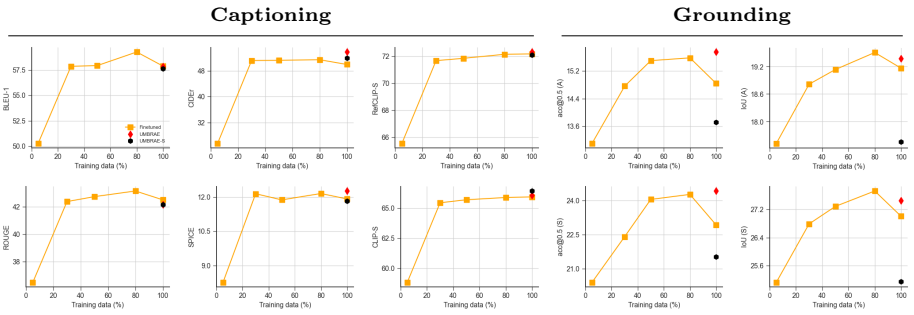
There are different sampling strategies of subjects and data samples in the cross-subject training. Our batch of B samples is made of $\theta \times B$ samples from the *same user* chosen according to users frequencies, while the remaining samples are then sampled from *other users*. This is illustrated on the left of Tab. 8 using, for simplicity, four users, $\theta = 0.44$ and $B = 16$. 'Random' means all subjects are randomly sampled, while 'Stratified' ensures that data samples from the four subjects are equal in number within a batch. 'Ours-R' and 'Ours' are the same for the dominant subject but differ in the sampling strategies for the remaining three subjects. Using a dominant subject per batch helps the model to learn intra-subject variations while being exposed to other subjects patterns to enhance inter-subject discrimination and alleviate catastrophic forgetting. Tab. 8 reports average metrics across users for 'Random', 'Stratified', 'Ours-R', and 'Ours' (using $\theta = 0.50$, $B = 256$). Ours outperforms all other sampling strategies.

4.4 Joint Grounding-Decoding Evaluation

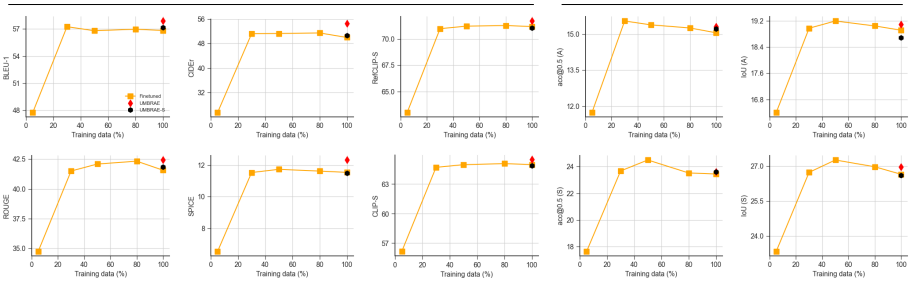
We visualize grounding results on the reference images (ground truth) to better assess their performance. Fig. 11 further shows grounding and reconstruction simultaneously to highlight their synergy. Results demonstrate that the two tasks are correlated. In some cases, although grounding is correct, reconstruction is inaccurate (*e.g.* surfer, giraffe, and skier are well located but misorriented). The first row presents reference images. The second row displays reconstructed images, which are generated using the decoded texts and groundings from the third row as inputs. The third row illustrates the spotting captioning results, where the coordinates for each mentioned object are omitted and instead visualized in color within the corresponding generated images shown in the second row.

4.5 Other Supported Tasks

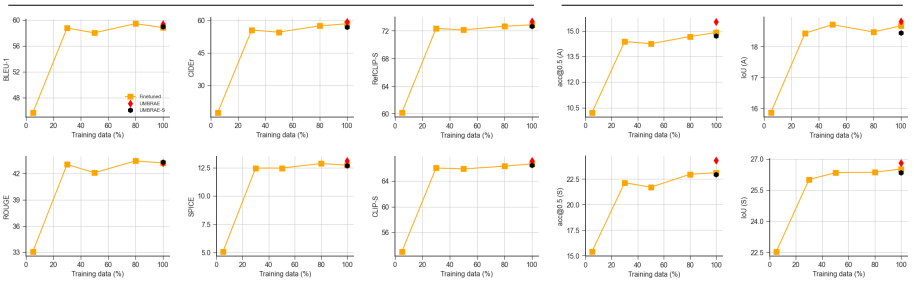
As we build on MLLM, we can explore a large variety of tasks. Tab. 1 lists the supported tasks, which can be categorized into three groups: captioning, grounding, and QA. We have presented the brain captioning results in Sec. 3.1 and the brain grounding (both REC and Spotting Captioning) results in Sec. 3.2. This section presents the additional QA tasks, including Q→A, Q→CA, and Q→C^{Box}A. Example results are shown in Fig. 12.



(a) Adaptation to S1 (trained on S2, S5, and S7)

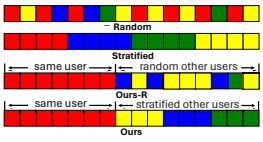


(b) Adaptation to S2 (trained on S1, S5, and S7)



(c) Adaptation to S5 (trained on S1, S2, and S7)

Fig. 10: Weakly-Supervised Subject Adaptation. We plot our performance for adaptation to a given subject (S1 in a, S2 in b, S5 in c). For adapting to a given subject Sx, we first train UMBRAE on all other subjects and then finetune the pretrained model with varying ratios of training data from the target subject.



Sampling	Captioning				Grounding			
	BLEU1	CIDEr	METER	RefCLIPS	acc@0.5 (A)	IoU (A)	acc@0.5 (S)	IoU (S)
Random	51.84	35.37	15.52	67.26	13.49	17.39	21.56	25.39
Stratified	58.91	55.83	18.94	72.69	17.31	20.34	27.08	29.03
Ours-R	58.02	55.02	18.68	72.09	15.22	18.84	23.58	26.74
Ours	**59.09**	**57.76**	**19.24**	**72.96**	**18.03**	**20.63**	**28.22**	**29.32**

Table 8: Comparison of Sampling Strategies. 'Random' means all subjects are randomly sampled, while 'Stratified' ensures that data from the four subjects are equal in number within a batch. 'Ours-R' and 'Ours' are the same when sampling from the selected dominant subject but differ for the remaining three subjects.

Fig. 11: Joint Visualization of Grounding and Reconstruction. The first row presents reference images. The second row displays reconstructed images generated using decoded texts and groundings from the third row. The coordinates in the third row for each mentioned object are omitted and visualized in color in the second row. Reconstruction may be inaccurate even with correct bounding boxes.

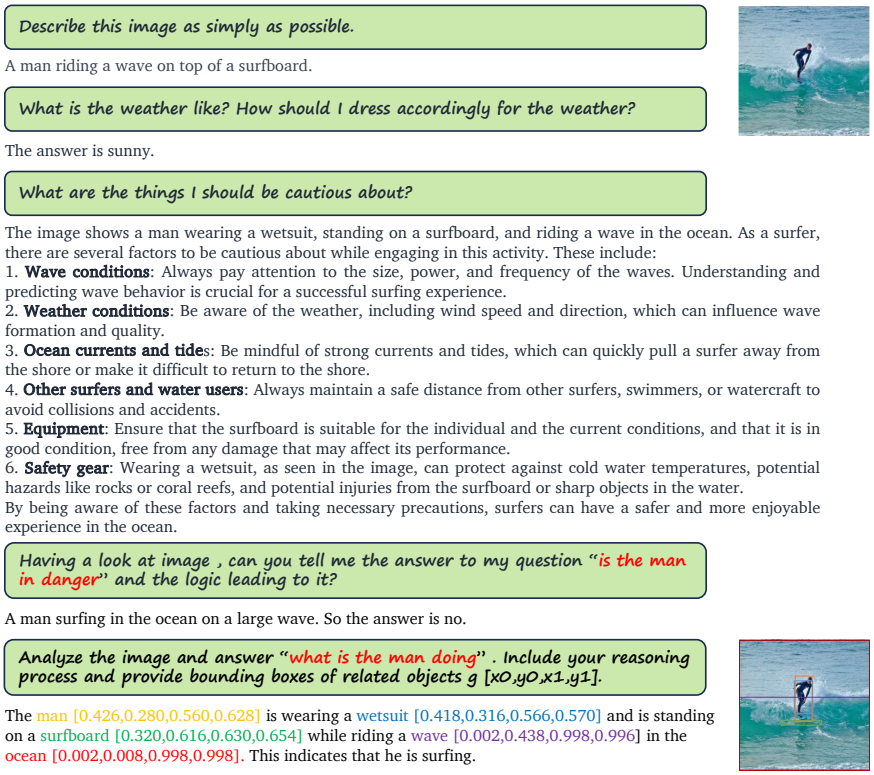


Fig. 12: Example Results of Other Tasks. The additional supported tasks listed in Tab. 1 include Q→A, Q→CA, and Q→C^{Box}A.

4.6 Model-Agnostic Performance

We employ the pretrained adapter and LLM from Shikra [4] due to its simultaneous support for both captioning and grounding. However, UMBRAE being model-agnostic, it can accommodate any image encoders (*e.g.*, CLIP [18] and DINO [3]), LLMs (*e.g.*, LLaMA [23] and Vicuna [5]), and MLLMs (*e.g.*, Shikra [4] and LLaVA [13]) according to specific needs.

To demonstrate this capacity, in Fig. 13, we provide qualitative results of 'UMBRAE with LLaVA' which replaces Shikra with the nonetheless popular LLaVA [13] as the MLLM. These example responses show that LLaVA [13] exhibits superior instruction-following capabilities in conversation, detailed description, and complex reasoning compared to Shikra [4]. However, LLaVA [13] lacks the capability of spatial localization.

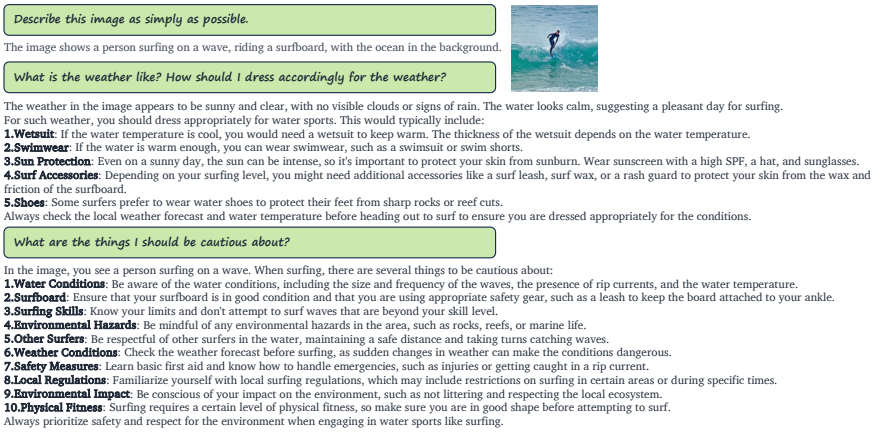


Fig. 13: UMBRAE with LLaVA as the MLLM. Here, we utilize UMBRAE though, replacing our default use of Shikra [4], with LLaVA [13] as the MLLM. The responses show that LLaVA exhibits superior instruction-following capabilities in conversation, detailed description, and complex reasoning compared to Shikra (The corresponding results from Shikra are in Fig. 12).

5 Discussion

5.1 Method Limitations

Our goal is to decode brain signals into multimodal explanations, with a primary emphasis on semantics and positions, achieving both concept recognition and spatial localization. While results demonstrate decent performance, there are instances where it falls short. Apart from previously mentioned challenges with inconspicuous objects, other failure cases can be categorized into three types:

(a) correct semantic recognition but inaccurate spatial localization, (b) accurate spatial localization but semantic errors, and (c) errors in both semantic recognition and spatial localization. We use spotting captioning to illustrate failure cases in Fig. 14, as this task outputs both concepts and locations.

Moreover, our method relies on pretrained MLLMs, inheriting their advantages while also facing common shortcomings associated with large models, including biases, hallucinations, generation of inappropriate content, and potential ethical concerns. Our method is also constrained by the quality of the captured brain responses in NSD [1] in two ways. Firstly, there are inherent inaccuracies introduced during data collection. NSD is captured using non-invasive neuroimaging techniques, where participants' compliance is necessary to avoid disruption in decoding caused by head movement or distraction. Secondly, the experimental images are sourced from COCO [12], which limits our method to natural scenes similar to those found in the COCO dataset [12].

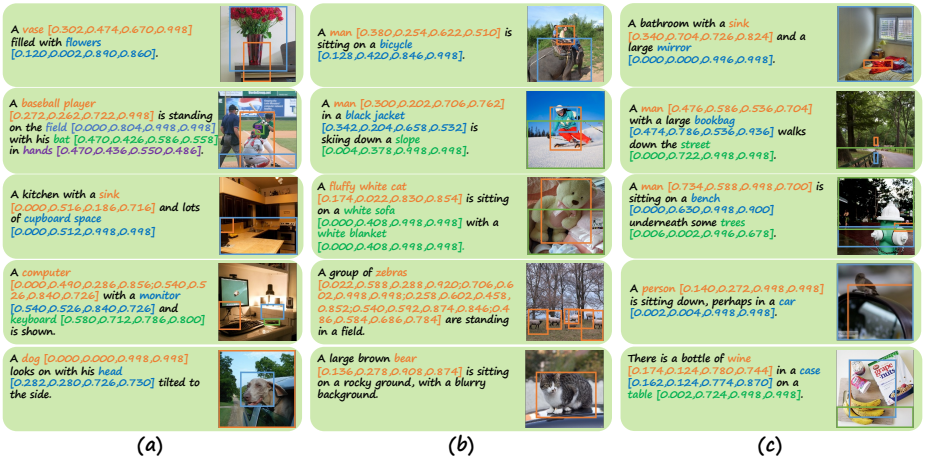


Fig. 14: Method Limitation. The failure cases can be categorized into: (a) correct semantic recognition but inaccurate spatial localization; (b) accurate spatial localization but semantic errors; (c) errors in both semantic recognition and spatial localization.

5.2 Potential Negative Impact

Our method relies on pretrained models as its foundation. While benefiting from the remarkable capabilities provided by LLMs [4, 7, 23], they also pose challenges and concerns that prompt broader societal impacts. These include potential biases in the training data, the generation of inaccurate or inappropriate content, and ethical considerations associated with their utilization. The inaccurate interpretation from our method may also lead to misunderstandings about the information contained within brain signals.

References

1. 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)
2. Arkhipkin, V., Filatov, A., Vasilev, V., Maltseva, A., Azizov, S., Pavlov, I., Agafonova, J., Kuznetsov, A., Dimitrov, D.: Kandinsky 3.0 technical report. arXiv preprint arXiv:2312.03511 (2023)
3. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV. pp. 9650–9660 (2021)
4. 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)
5. 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)
6. 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)
7. 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)
8. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. In: EMNLP (2021)
9. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. NeurIPS **30** (2017)
10. 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)
11. Lin, S., Sprague, T., Singh, A.K.: Mind Reader: Reconstructing complex images from brain activities. NeurIPS **35**, 29624–29636 (2022)
12. 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)
13. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. NeurIPS **36** (2023)
14. 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)
15. Ozcelik, F., VanRullen, R.: Brain-Diffuser: Natural scene reconstruction from fMRI signals using generative latent diffusion. Scientific Reports **13**(1), 15666 (2023)
16. Parmar, G., Zhang, R., Zhu, J.Y.: On aliased resizing and surprising subtleties in gan evaluation. In: CVPR. pp. 11410–11420 (2022)
17. 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)
18. 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)
19. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)

20. 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)
21. Takagi, Y., Nishimoto, S.: High-resolution image reconstruction with latent diffusion models from human brain activity. In: CVPR. pp. 14453–14463 (2023)
22. 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)
23. 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)
24. Uddin, L.Q.: Salience processing and insular cortical function and dysfunction. Nature reviews neuroscience **16**(1), 55–61 (2015)
25. Xia, W., de Charette, R., Öztireli, C., Xue, J.H.: Dream: Visual decoding from reversing human visual system. In: WACV. pp. 8226–8235 (2024)
26. 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)
27. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR. pp. 586–595 (2018)