UMBRAE: Unified Multimodal Brain Decoding (ECCV 2024)

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

1 UCL

2 Inria INVENTEURS DU MONDE NUMÉRIQUE

3 UNIVERSITY OF CAMBRIDGE

Problem Statement

Visual stimuli

Brain Activities

fMRI voxels

Reconstruction

Neural (Visual) Decoding

Motivation:

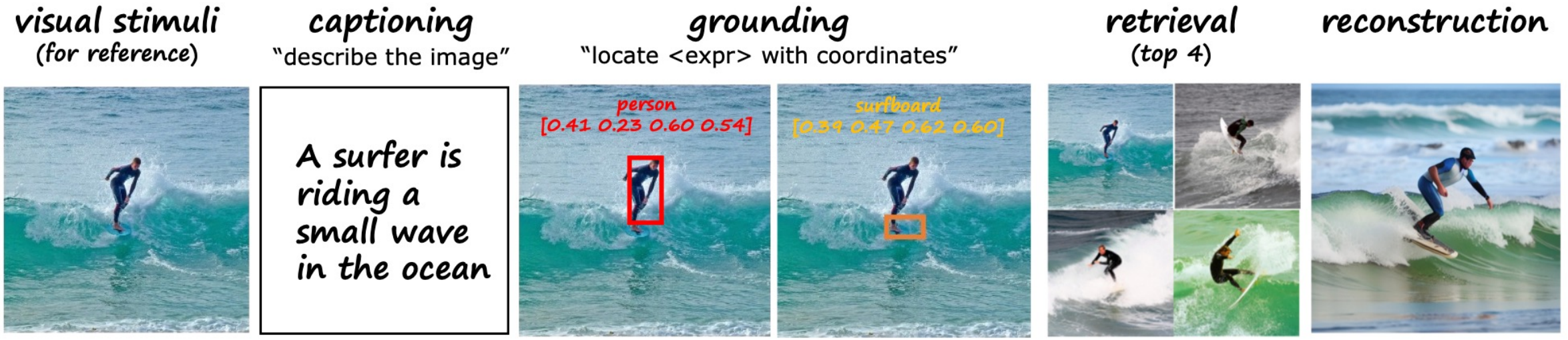
Brain signal is indirectly understandable by humans.

- Can we interpret brain signals?

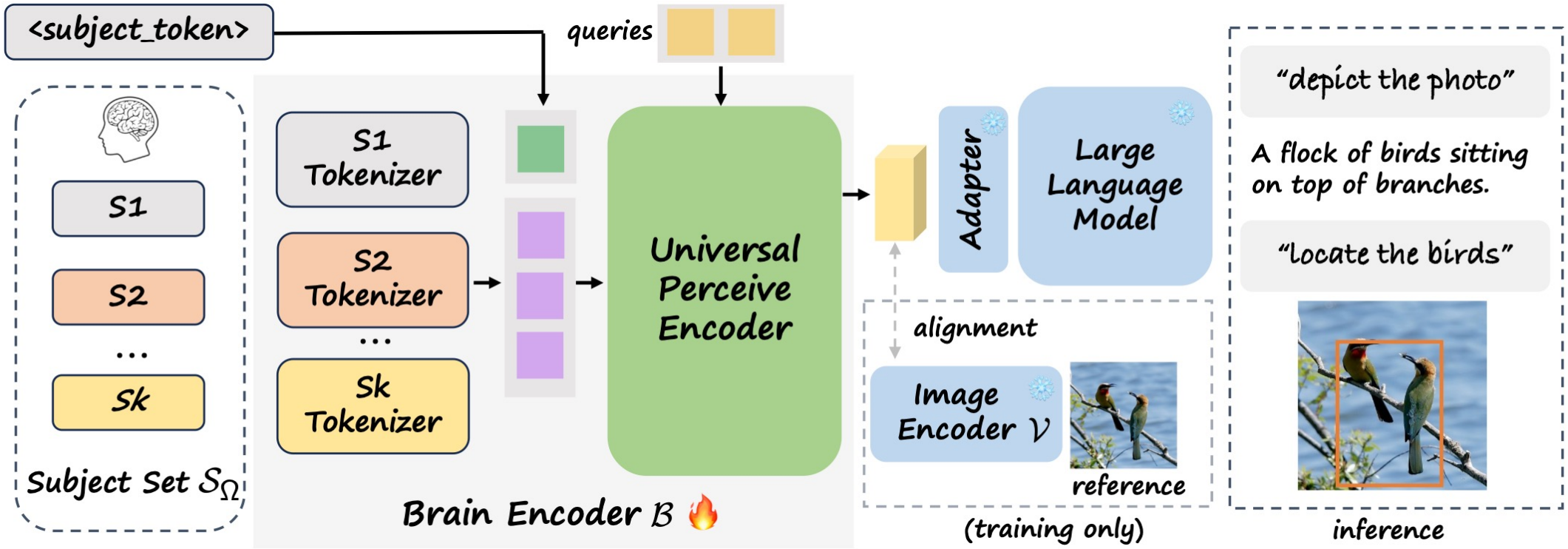
Brain patterns across individuals are unique.

- Can we train a model for all subjects?

UMBRAE performs Unified Cross-Subject Multimodal Brain Decoding using pretrained MLLMs with prompts.



Framework



Overview of UMBRAE Framework. The **brain encoder** includes (a) subject-specific tokenizers and (b) a universal perceive encoder. Brain signals from multiple subjects are mapped into a common feature space, enabling **cross-subject training** and **weakly-supervised subject adaptation**. The brain encoder is trained to align neural signals with image features only, without the need for captions or bounding boxes during the training process.

During inference, the learned brain encoder interacts with MLLMs and performs a variety of brain understanding tasks at different levels of granularity according to given prompts as the instructions.

Demo Examples

fMRI input

visual stimuli
(reference only)

"Describe this <image> as simply as possible."

A person holding a tennis racket in his hands.

"What is he wearing?"

He is wearing a white shirt and dark shorts.

"Describe this <image> as simply as possible."

A man riding skis down the side of a snow covered slope.

"How is the weather in the image?."

The weather appears to be sunny and clear.

(a) Brain Captioning

(Describe brain signals using texts)

"Please interpret this image and give coordinates [x1,y1,x2,y2] for each object you mention."

The image portrays a man [0.288,0.154,0.714,0.998] in a white shirt playing tennis in a grassy field. He appears to be swinging a racket at a ball.

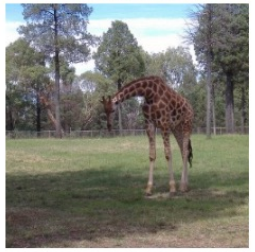
It depicts a photo of person [0.300,0.238,0.738,0.808] skiing down a snowy slope [0.004,0.440,0.998,0.998]

(b) Brain Grounding

(Locate visual concepts using boxes)

Brain Captioning

Comparison with SOTAS. We are the only method that does not require captions during training.

Reference	*Captioning*	
	SDRecon	the sea with some trees in the fore, and mountains in the distance are red
	BrainCap	a person is standing on a beach with a snowboard.
	OneLLM	A group of people gathered on the beach flying kites.
	UMBRAE-S1	A group of people riding boards on top of a beach.
	UMBRAE	A person is parasailing on a lake with mountains in the distance.
	SDRecon	the city of london from an perspective
	BrainCap	a corner of a building with a train station.
	OneLLM	A kitchen is seen through an open door.
	UMBRAE-S1	A large building with a clock tower on top.
	UMBRAE	A large building with a clock tower on top.
	SDRecon	some animals in the wild area near to wildlife world
	BrainCap	a large area of grass.
	OneLLM	A man standing on a snowy slope skiing.
	UMBRAE-S1	A giraffe is standing in a grassy field.
	UMBRAE	A giraffe is standing in a grassy field.
	SDRecon	a beach with rocks on the shore and ocean
	BrainCap	a large body of water with some birds on it.
	OneLLM	A foggy day in San Francisco with a bus and a streetlamp.
	UMBRAE-S1	A group of birds standing on top of a sandy beach.
	UMBRAE	A flock of birds standing on a body of water.

UMBRAE-Sx means model trained with single-subject data; UMBRAE is a cross-subject model;
Reference images are visual stimuli for input brain responses and are just used here for visualization.

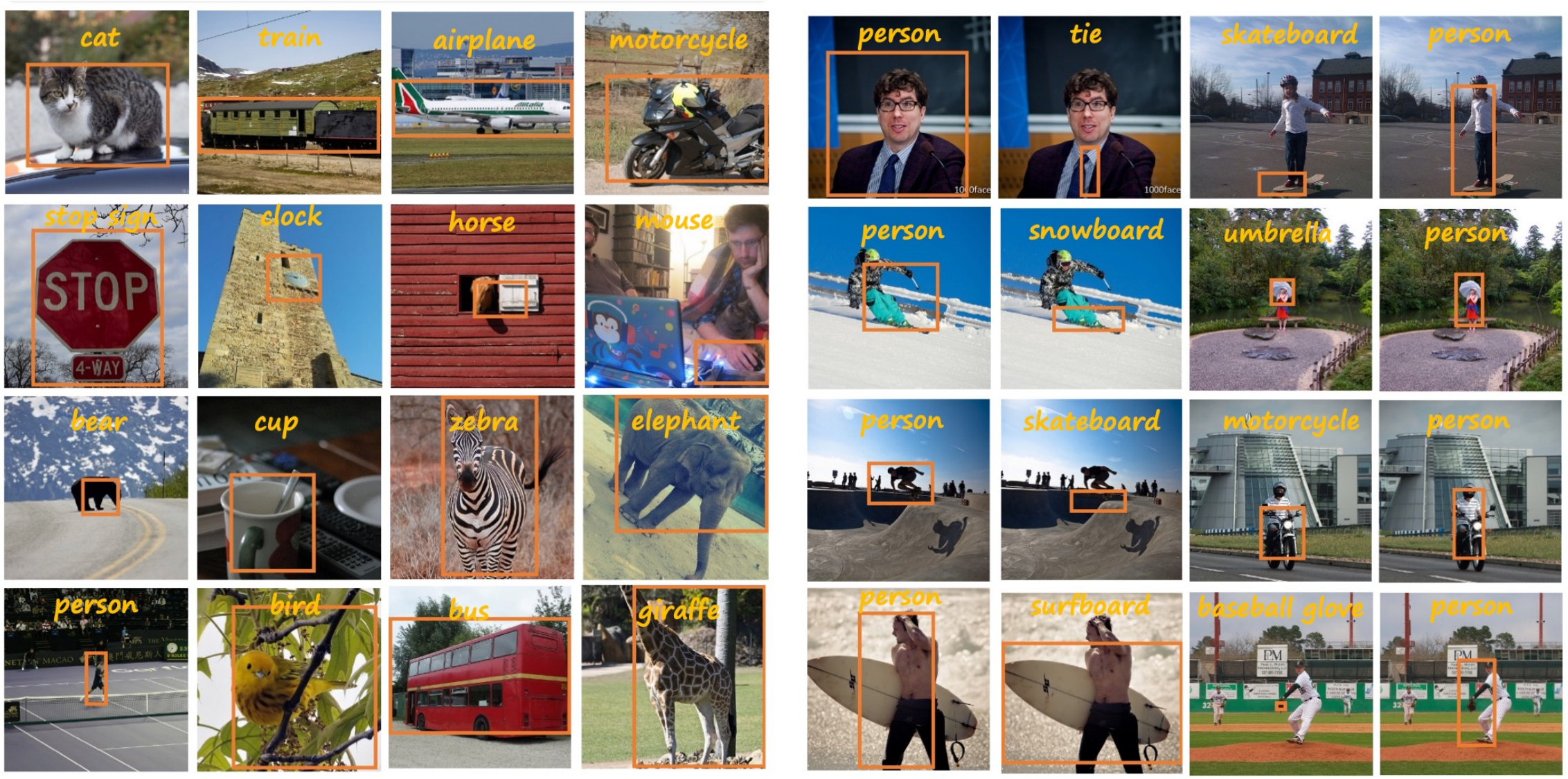
Brain Captioning

Results on Different Subjects

Reference	*Captioning*	
	COCO	A bathroom with a vanity mirror sitting above a toilet next to a bathtub.
	Shikra-w/img	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	A picture of a cat and some luggage.
	Shikra-w/img	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	A large field of grass with sheep grazing on the land.
	Shikra-w/img	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	A man riding a snowboard down a hill.
	Shikra-w/img	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.

Brain Grounding

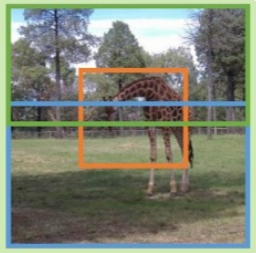
Referring Expression Comprehension: “Locate <expr> in <image> and provide its coordinates, please”



Brain Grounding

Spotting Captioning: “Please interpret this image and give coordinates [x1,y1,x2,y2] for each object you mention.”

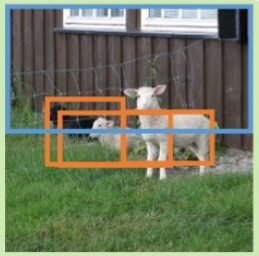
A giraffe [0.300,0.262,0.746,0.670] is standing in a grassy field [0.002,0.400,0.998,0.998] with trees [0.000,0.000,0.998,0.498] in the background.



A man [0.306,0.082,0.682,0.578] in a hat [0.432,0.054,0.592,0.162] rides a horse [0.278,0.158,0.782,0.998].

A closeup of a fire hydrant [0.382,0.082,0.672,0.792].

Three sheep [0.470,0.426,0.840,0.652;0.162,0.378,0.482,0.654;0.212,0.432,0.672,0.658] are standing near a fence [0.000,0.000,0.998,0.520] .

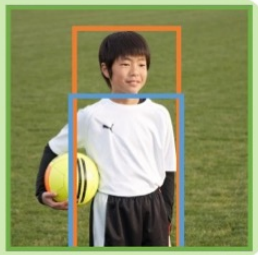


Two zebras [0.014,0.344,0.438,0.544;0.470,0.330,0.996,0.624] are standing in a field [0.000,0.548,0.998,0.998] with trees [0.000,0.002,0.998,0.402] in the background.

A large clock [0.320,0.000,0.738,0.528] on the side of a building [0.004,0.004,0.998,0.998].

A person [0.350,0.344,0.664,0.586] is skateboarding on a ramp [0.000,0.544,0.998,0.998].

A boy [0.270,0.088,0.708,0.998] in a white shirt [0.256,0.362,0.720,0.998] is standing on a grassy field [0.000,0.000,0.998,0.998].



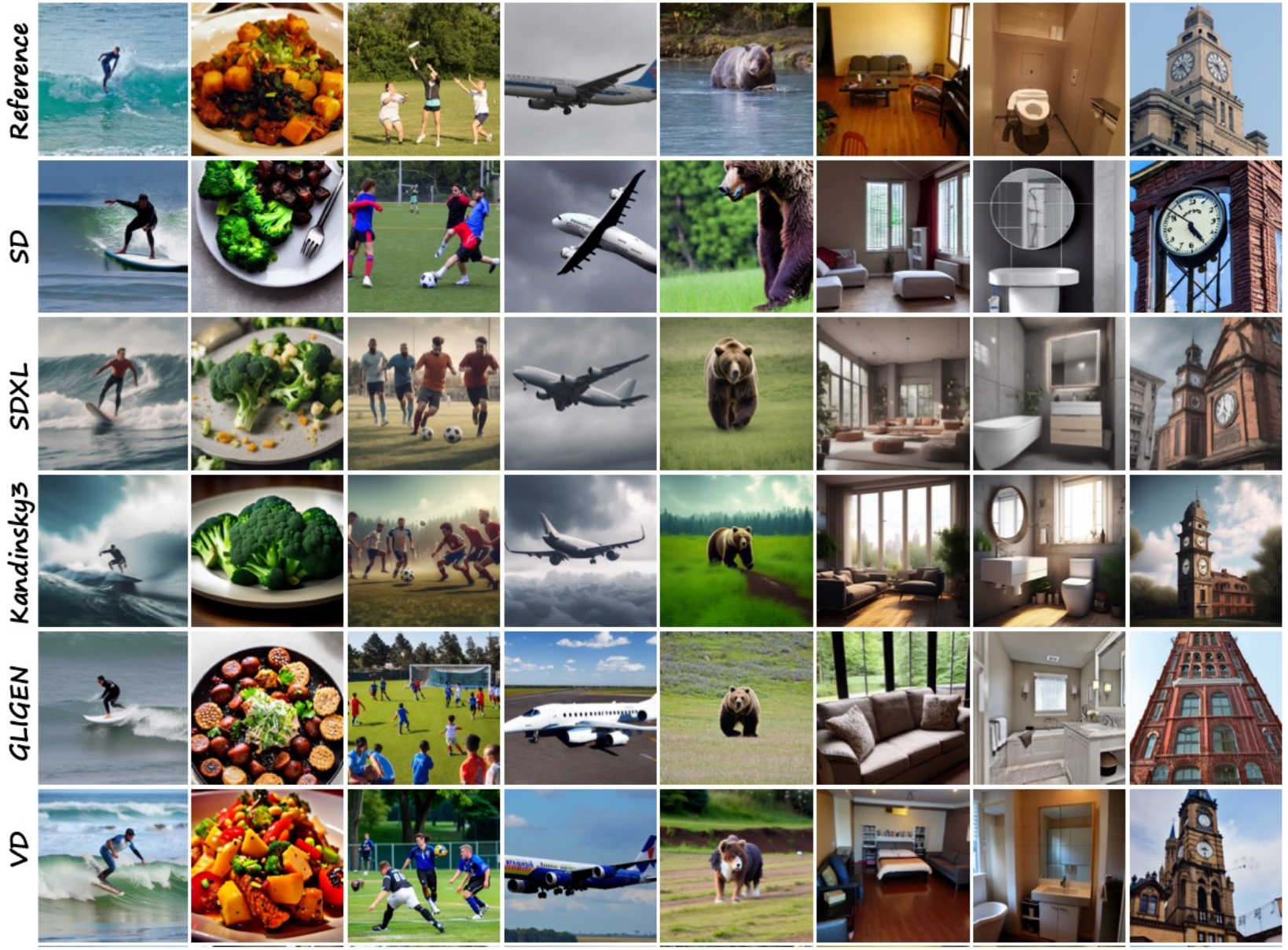
Brain Grounding

Referring Expression Comprehension

"Locate <expr> in <image> and provide its coordinates, please"

Results on Different Subjects

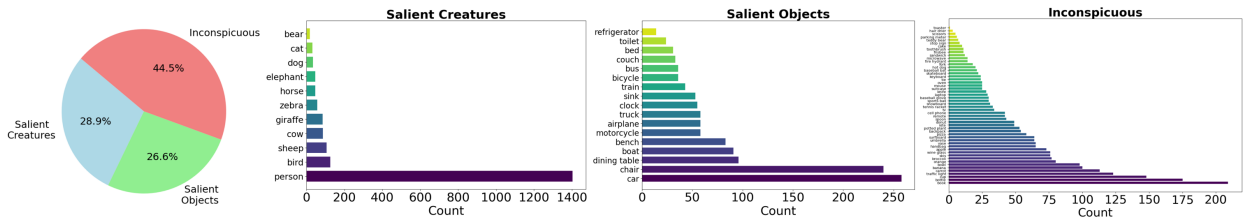
Visual Decoding



BrainHub

For evaluation, we introduce BrainHub, a brain understanding benchmark, based on NSD and COCO.

There are 982 test images, 80 classes, 4,913 captions, and 5,829 boundingboxes. For grounding evaluation, we further group the 80 classes of COCO into 4 salience categories according to their salience in images: Salient (S), Salient Creatures (SC), Salient Objects (SO), and Inconspicuous (I). The illustration shows the statistics and mapping of our categories, w.r.t. COCO classes.



(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.

Quantitative Comparison

brain captioning

Method	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE	CIDEr	SPICE	CLIP-S	RefCLIP-S
Shikra-w/img [9]	82.38	69.90	58.63	49.66	35.60	65.49	161.43	27.62	80.60	85.92
SDRecon [44]	36.21	17.11	7.72	3.43	10.03	25.13	13.83	5.02	61.07	66.36
OneLLM [16]	47.04	26.97	15.49	9.51	13.55	35.05	22.99	6.26	54.80	61.28
UniBrain [32]	-	-	-	-	16.90	22.20	-	-	-	-
BrainCap [14]	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

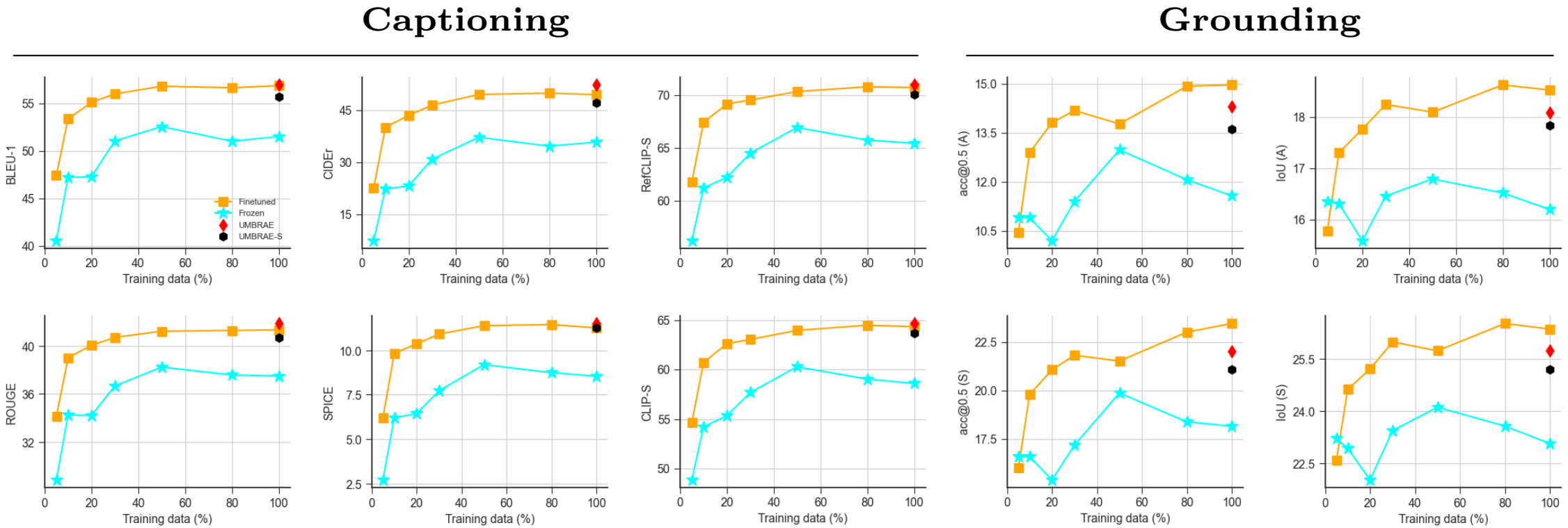
brain grounding

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 [9]	*	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 [34]	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 [41]		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 [53]		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.

Weakly-Supervised Adaptation

Adapting a new subject to a pretrained brain encoder, using only a portion of training examples: this model for S7 is trained or finetuned on a pretrained model (trained on S1, S2, and S5) using varying ratios of training data



Ablation Studies: network architecture

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

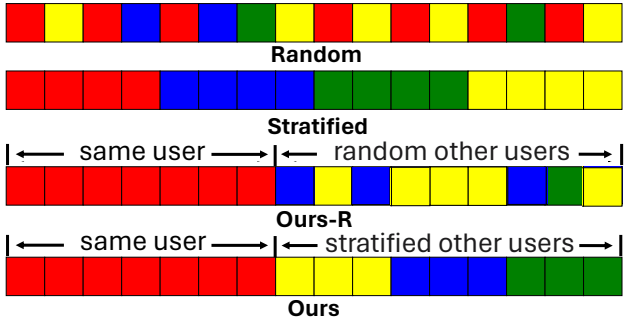
MLP: MLP-based brain encoder

Enc-S and **Enc-U**: the single and cross-subject encoders

The adapter operates under three distinct training settings: **Pretrained**, **Finetuned**, and **Joint**, with different **Loss Types** applied to the outputs with **Dim.** from either the encoder (**E**) or the adapter (**A**)

Ablation Studies: sampling strategy in cross-subject training

Comparison (right) of using different **sampling strategies** (left) when training the cross-subject brain encoder



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**

Random: all subjects are randomly sampled

Stratified: 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 remaining three subjects.

Failure Cases

The performance relies on the salience of objects in the image, especially suffers when dealing 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.

Thank you for you attention

Paper

Project

Code

Email: weihao.xia.21@ucl.ac.uk