ICML
International Conference
On Machine Learning

Code

Mind-Omni: A Unified Multi-Task Framework for Brain-Vision-Language Modeling via Discrete Diffusion

• Yizhuo Lu[1,2], Changde Du*[1,3,4], Qingyu Shi[5], Hang Chen[1,2], Peng Jie[1,3], Liuyun Jiang[2], Shuangchen Zhao[1,3,4], Huiguang He[Ŧ1,2,3,4],

• [1] NeuBCI Lab, State Key Laboratory of Brain Cognition and Brain-inspired Intelligence Technology, CASIA,

• [2] School of Future Technology, University of Chinese Academy of Sciences, [3] School of Artificial Intelligence, UCAS,

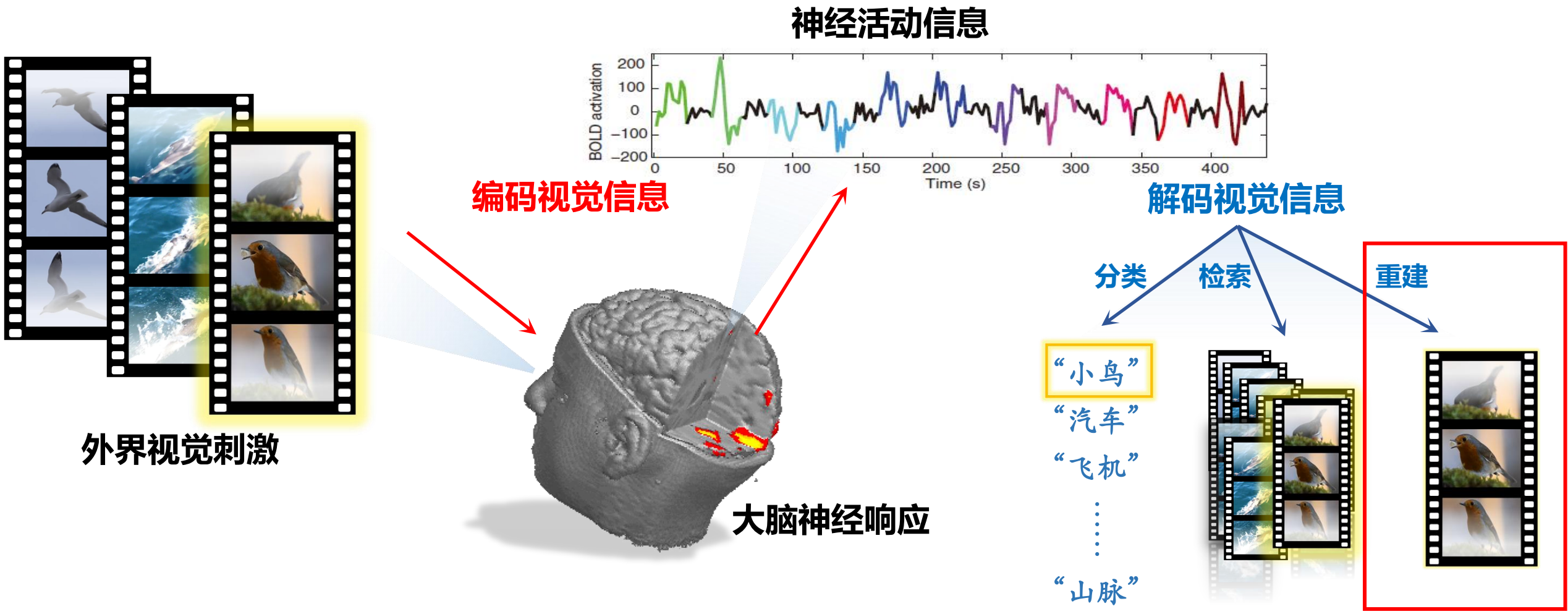
• [4] Zhongguancun Academy, [5] Peking University

• *Email: luyizhuo2023@ia.ac.cn, huiguang.he@ia.ac.cn*

1 中国科学院自动化研究所 INSTITUTE OF AUTOMATION CHINESE ACADEMY OF SCIENCES

2 中国科学院大学 University of Chinese Academy of Sciences

背景介绍（神经编解码）

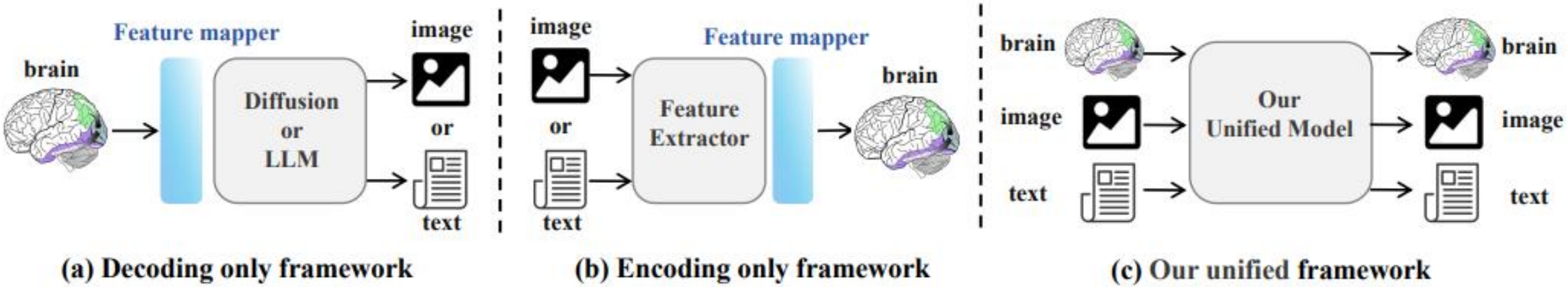


视觉神经信息编解码是建模**外部视觉刺激**和**内部神经活动表征**两者之间关系的主要研究手段。

相关工作

任务专属架构：传统范式为神经编码和解码任务分别设计不同的架构，模型参数效率低下

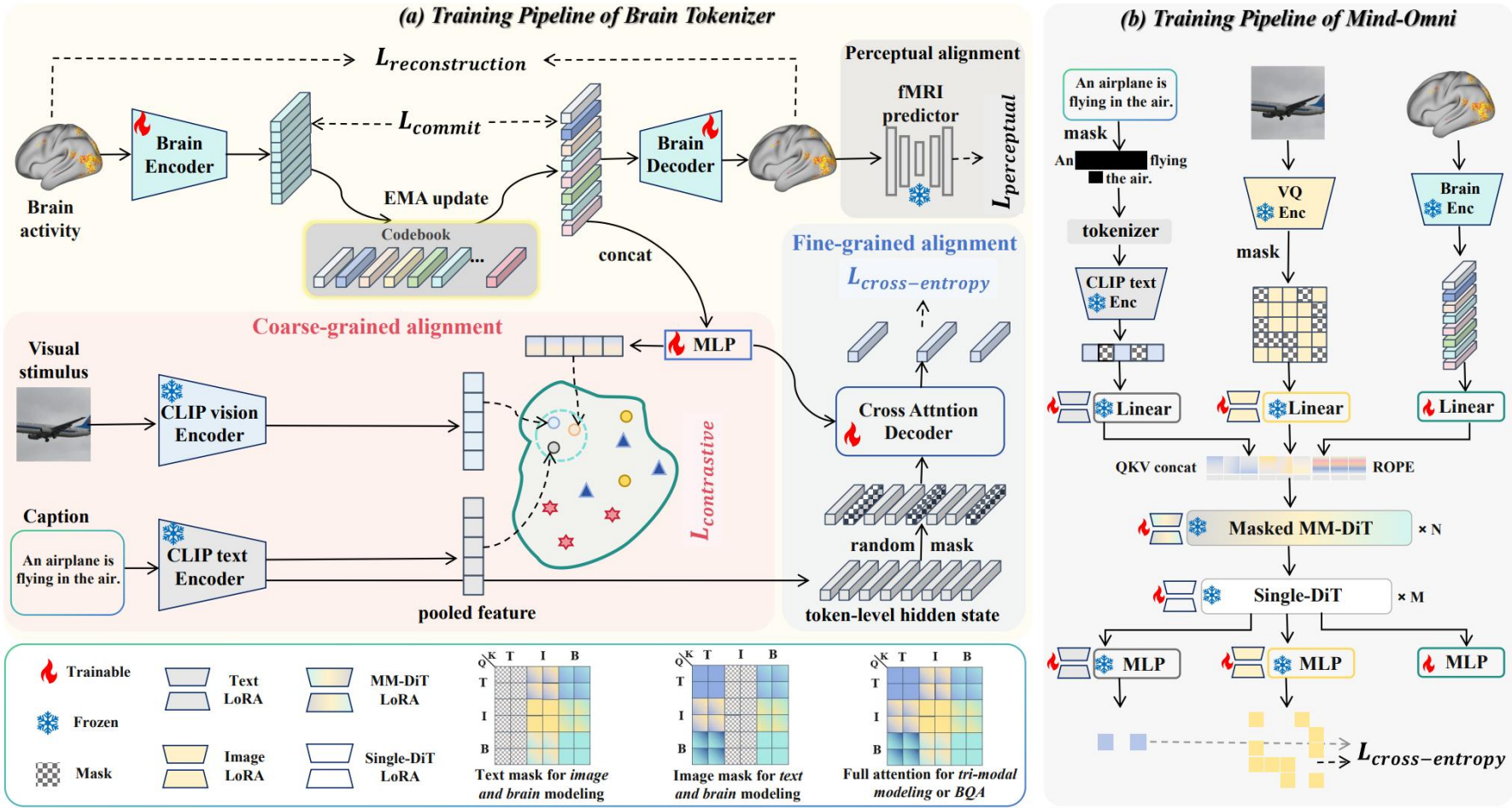
通用架构：统一的编解码基座模型，实现了任意模态的交互



Method	Subject	Encoding			Decoding			
		I→B	T→B	I&T→B	B→I	B→T	B→I&T	BQA
MindEye[NeurIPS2024] [66]	Single				✓			
MindBridge[CVPR2024] [86]	Multi				✓			
Psychometry[CVPR2024] [61]	Multi				✓			
BrainCap[NeurIPS2023 W] [22]	Single				✓	✓		
UMBRAE[ECCV2024] [91]	Multi					✓		✓
Sem-Recons[Nat. Neurosci.] [80]	Single					✓		
CLIP-Map[Nat. Mach. Intell.] [84]	Single	✓						
MindSimulator[ICLR2025] [5]	Single	✓						
BraVL[TPAMI2023] [13]	Single	✓	✓	✓	✓			
Ours	Multi	✓	✓	✓	✓	✓	✓	✓

我们利用离散扩散模型建模了神经编解码中的7种任务，包括：图像重建，语言重建，视觉问答等

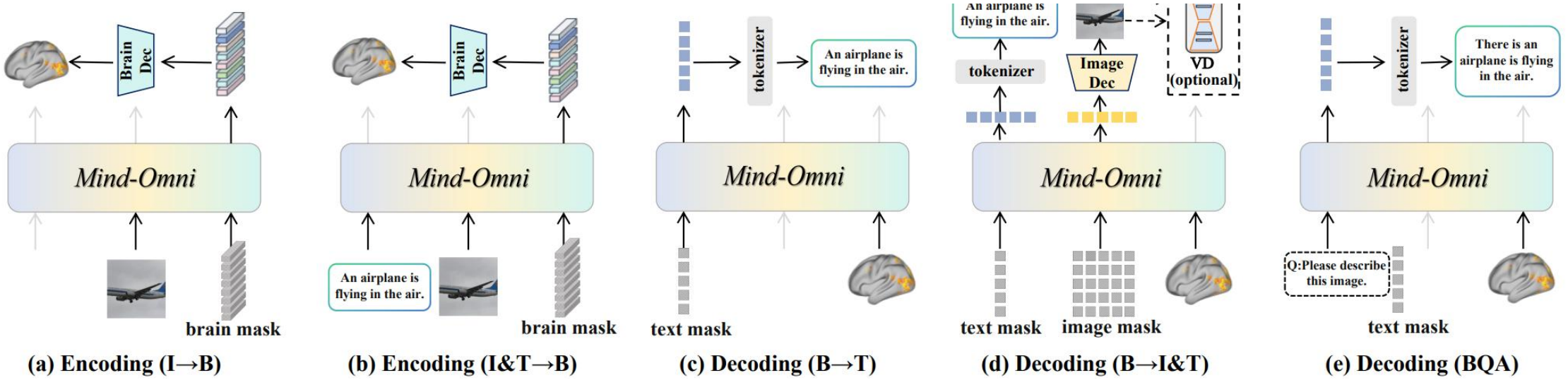
研究方法（训练框架）



1.训练带语义监督的Brain Tokenizer把连续的脑信号转换成离散token

2.适配MM-DiT以接受第三个模态的输入

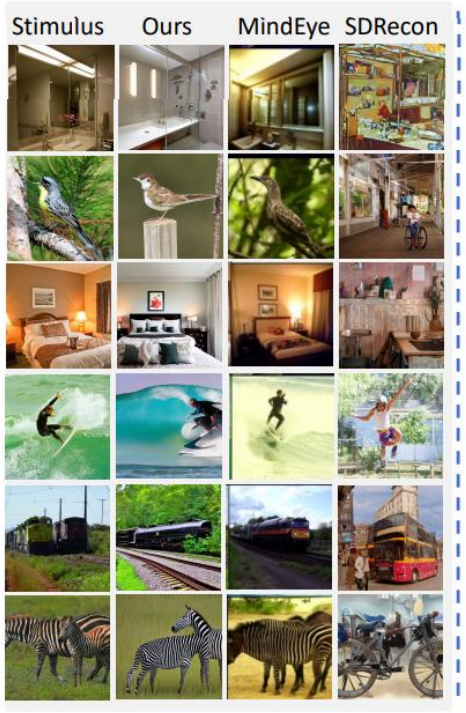
研究方法（推理框架）



在推理阶段输入干净的条件模态，待生成模态由随机噪声初始化。

实验结果

视觉重建



自然语言问答

Stimulus (reference only)

Brief Description

Q: Give me a very short description of the scene.

A: a train was traveling on the track.

Detail Description

Q: Write a detailed description of the given image.

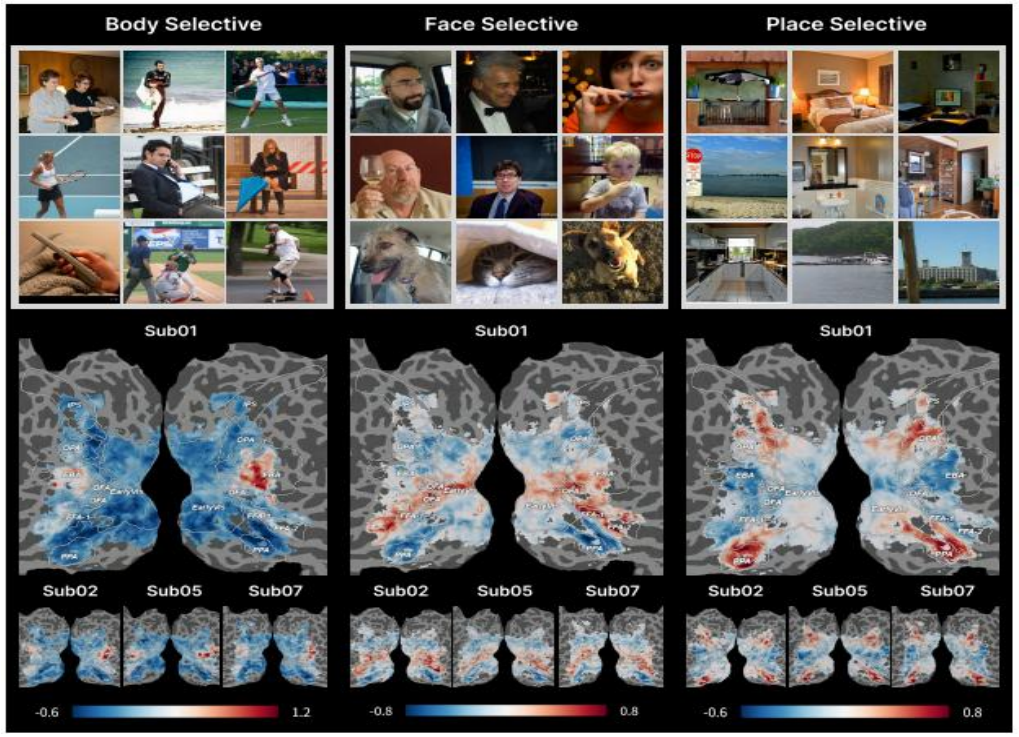
A: two vintage passenger trains, painted in black and yellow accents, travels along a track, surrounded by lush green trees under a clear blue sky.

Reasoning

Q: What is the status of the two trains in the picture?

A: two passenger trains are traveling side by side on parallel tracks, as they can be seen serving as a convenient travel for their respective destinations.

神经编码



编解码基座模型统一了神经编码和神经解码领域中的7种不同的任务

模型成功习得了视皮层的内在功能组织规律，自发涌现出了与人脑一致的类别选择性（如 FFA/PPA）

实验结果

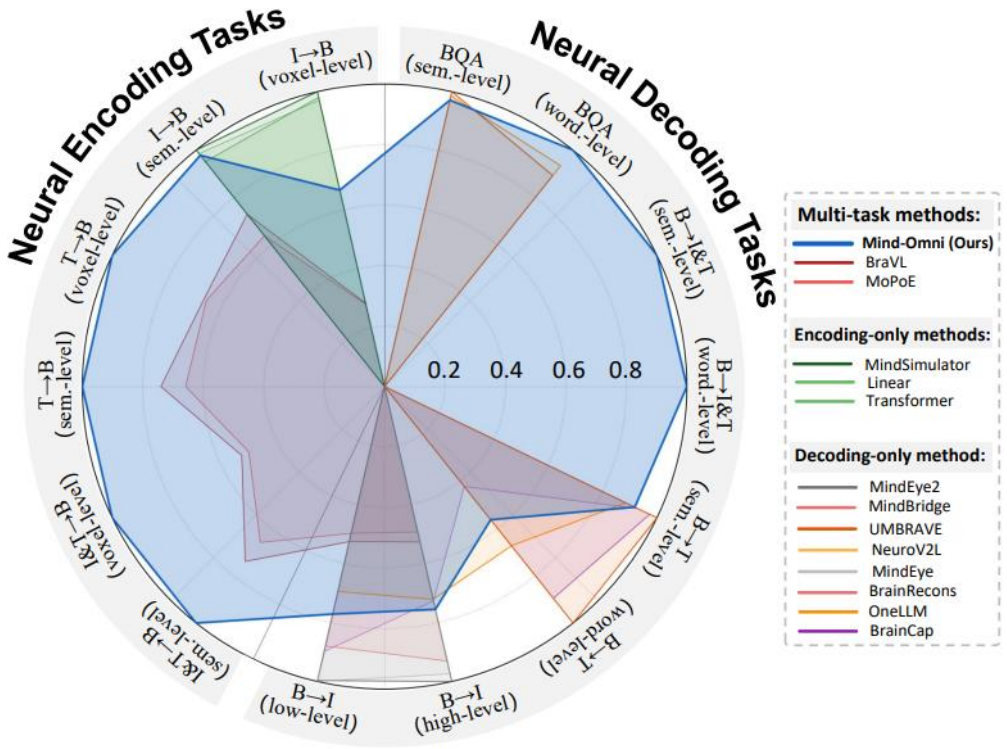


Table 2. Quantitative evaluation of visual reconstruction (B→I and B→I&T tasks) averaged over subjects 1, 2, 5, and 7. The best, and second-best results among comparable models are highlighted in **bold**, and with an underline, respectively.

Method	Trainable Parameters	# Models	# Tasks	Low-Level				High-Level			
				PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	EffNet-B ↓	SwAV ↓
SDRecon [77]	3B	4	2	–	–	**83.0%**	83.0%	76.0%	77.0%	–	–
MoPoE [76]	564M	4	4	.021	.145	54.4%	56.2%	59.3%	57.6%	.965	.752
BraVL [13]	564M	4	4	.023	.167	56.3%	59.1%	61.7%	63.5%	.943	.757
OneLLM [27]	7B	4	7	.053	.313	64.7%	76.1%	75.3%	77.2%	.851	.551
Ours (B→I)	442M	1	7	**.118**	**.383**	67.1%	72.8%	69.4%	66.7%	.918	.583
Ours (B→T)	442M	1	7	.036	.284	62.8%	70.4%	67.9%	69.9%	.895	.623
Ours (B→I&T)	442M	1	7	.058	.341	72.5%	**84.9%**	**78.8%**	**79.8%**	**.824**	**.537**

Table 3. Quantitative analysis of detailed descriptions (B→T and B→I&T tasks), and reasoning (BQA task). The best and second-best results are indicated in **bold** and with an underline, respectively. LLM-as-Judge refers to evaluation using Qwen3-VL-30B-A3B-FP8: given the stimulus image, question, reference answer, and model output, the judge determines whether the answer is correct.

Method	Train. Params	External LLM	# Models	# Tasks	BLEU1 ↑	BLEU2 ↑	BLEU3 ↑	METEOR ↑	ROUGE ↑	CIDEr ↑	SPICE ↑	CLIP ↑	RefCLIP ↑	BERT ↑	LLM-as-Judge ↑
					Detail Description										
UMBRAE [91]	146.57M	Vicuna (13B) [10]	1	2	21.39	11.86	6.31	11.31	17.60	6.04	6.43	**60.85**	**65.96**	–	–
OneLLM [27]	7B	LLaMA2 (7B) [15]	4	7	18.41	12.13	5.34	9.37	17.13	9.41	6.34	50.31	51.43	86.18	–
Ours (B→T)	442M	None	1	7	14.92	9.03	5.83	13.86	13.35	6.28	6.78	48.47	47.00	80.21	–
Ours (B→I&T)	442M	None	1	7	**29.12**	**17.63**	**11.36**	**26.05**	**30.54**	**12.26**	**13.25**	53.67	52.75	**87.73**	–
					Reasoning										
UMBRAE [91]	146.57M	Vicuna (13B) [10]	1	2	**46.33**	**31.42**	**23.56**	41.93	42.12	156.81	33.70	69.57	75.22	91.33	**25.48**
OneLLM [27]	7B	LLaMA2 (7B) [15]	4	7	19.27	13.77	10.76	42.83	45.63	223.67	37.43	67.46	73.41	**91.93**	19.12
Ours (BQA)	442M	None	1	7	23.18	15.83	11.86	**50.13**	**52.91**	**223.98**	**43.28**	**70.65**	**76.72**	81.96	24.37

编解码统一模型在语义级任务上取得了和任务特定模型相近甚至更优的性能

但是在像素级（解码）和体素级（编码）任务上扔落后于任务特定模型

消融实验

Table 5. Ablation study on the Brain Tokenizer's architecture design, where rPCC is calculated on self-reconstructed fMRI signals. The chance level for retrieval is 0.05.

$\mathcal{L}_{SA}$	$\mathcal{L}_{perceptual}$	codebook size	code dim.	code num.	rPCC ↑	Retrieval (Top50) ↑		codebook usage ↑
						B2I	B2T	
×	×	64	512	64	0.37	0.05	0.05	1%
×	×	64	128	64	0.39	0.05	0.05	6%
×	×	64	16	64	0.43	0.05	0.05	70%
×	×	128	16	64	0.45	0.05	0.05	32%
✓	×	64	16	64	0.64	0.58 (+0.53)	0.54 (+0.49)	100% (+30%)
✓	×	128	16	64	0.68	0.60	0.57	62%
✓	×	128	32	32	0.64	0.61	0.58	38%
✓	×	256	16	64	0.63	0.60	0.57	40%
✓	×	384	16	64	0.64	0.60	0.56	35%
✓	×	512	16	64	0.62	0.58	0.53	28%
✓	✓	128	16	64	**0.68**	0.62 (+0.02)	0.59 (+0.02)	80% (+18%)
✓	✓	128	32	32	0.64	**0.68** (+0.07)	**0.64** (+0.06)	40% (+2%)

由于脑信号信噪比较低，因此训练tokenzier的时候需要小codebook以及较少的token。

Table 6. Ablation study on the impact of different training strategies. Blue-shaded row denote the strategy used in our model. Complete results are provided in Tabs. 13–18.

Task	Decoding			Encoding		
	BLEU1↑	ROUGE↑	BERT↑	gPCC↑	gMSE↓	gRSA↑
(a) Choice of Tokenization Strategy						
code dim.=32	29.56	**25.85**	**88.32**	0.126	**0.621**	0.315
code dim.=16	**30.28**	25.73	88.04	**0.145**	0.698	**0.342**
(b) Choice of Training Strategy						
Direct	29.11	27.31	84.29	0.132	0.677	0.369
Progressive	**29.12**	**30.54**	**87.73**	**0.160**	**0.654**	**0.408**
From Scratch	20.35	14.32	74.66	0.104	0.984	0.242
From Pretrained	**29.12**	**30.54**	**87.73**	**0.160**	**0.654**	**0.408**
(c) Choice of Image-Caption Pairs						
Raw COCO	24.37	26.52	83.17	0.133	0.671	0.384
Qwen2-VL Enhanced	**29.12**	**30.54**	**87.73**	**0.160**	**0.654**	**0.408**

高质量预训练与caption数据标注对于后训练性能影响明显

Code

Thanks!