

大语言模型与脑神经编解码方法

宗成庆

中国科学院自动化研究所



1. 引言
2. 数据集构建
3. 脑神经编解码
4. 结束语

1. 引言



2022年11月29日, OpenAI 公司发布 ChatGPT, 允许用户以自然语言的形式提问任何问题, 系统都可以给出较好的回答。

- 2023年3月14日, GPT-4.0
- 2024年2月15日, Sora
- 2024年5月1日, GPT-4o (Omni)
- 2024年9月12日, o1-preview, o1-mini
- 2024年12月20日, o3, o3-mini; 2025年1月31日 o3-mini 正式上线
- 2025年3月25日, OpenAI 宣布GPT-4o 增加图像生成功能
- 文心一言(百度), 豆包(字节), 通义千问(阿里), Kimi (月之暗面), 九天, 星火(讯飞), 腾讯混元, 百川, 华为盘古, 日日新(商汤), 紫冬太初, 智谱华章, DeepSeek

✧ **统一框架**

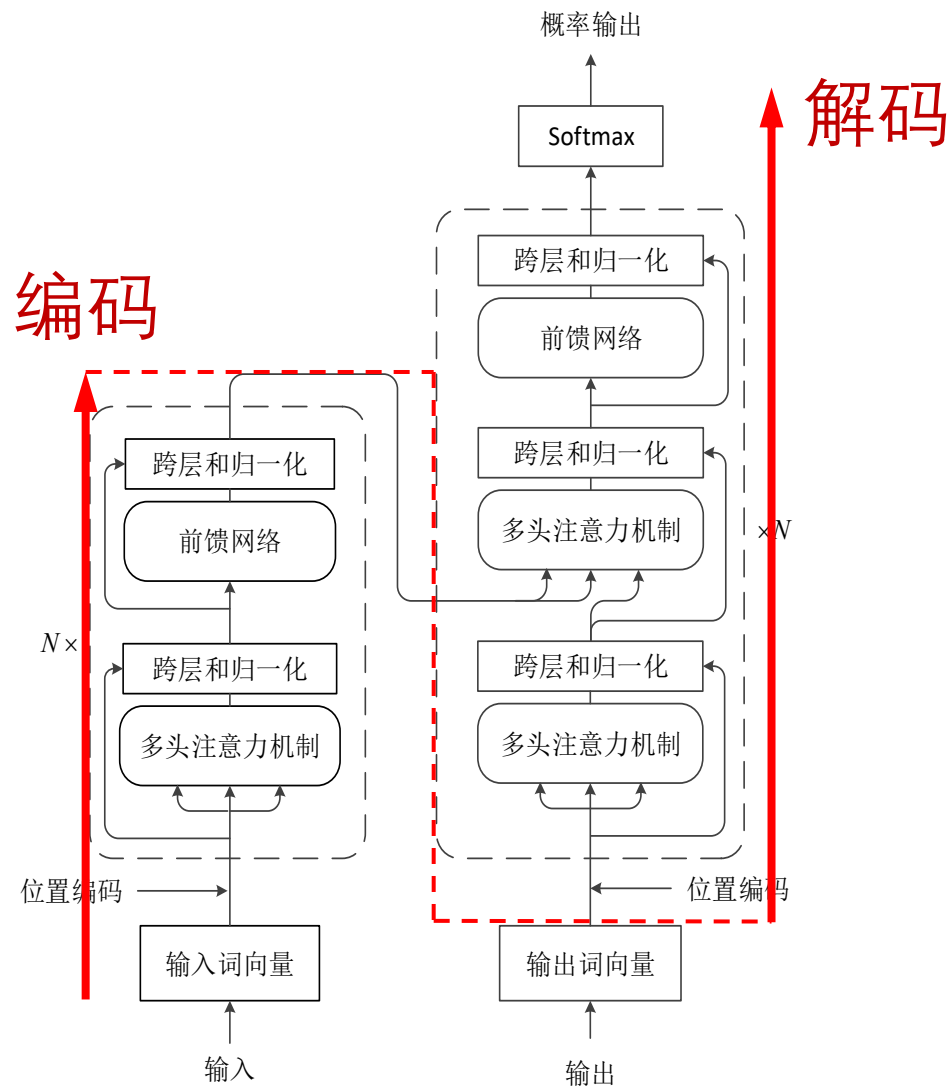
✧ **不同语言、不同模态信息、不同任务几乎无差别处理**

1. 引言

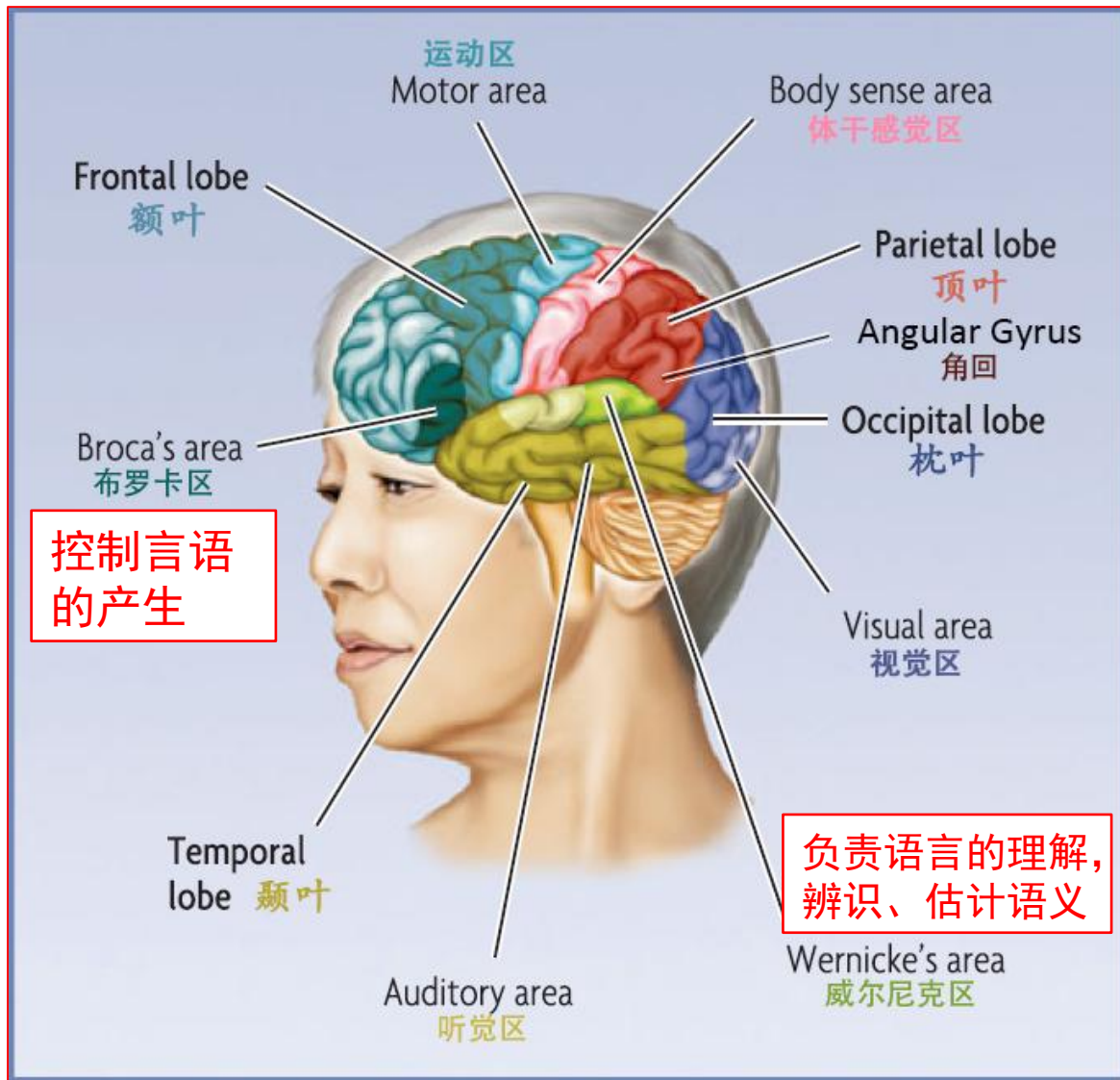
谷歌公司于2017年提出了基于自注意力机制的端到端 (End-to-End, E2E) 的转换框架，并将其命名为Transformer。

架构结构如右图所示。

- Machine translation
- Human-dialogue system / Q&A
- Sentiment/ emotion analysis
- Text classification
- Automatic summarization
- Automatic writing / language generation
- Speech recognition
- Image caption
-



1. 引言

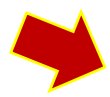


- 人脑是如何表征和记忆语言的？
- 能否根据神经影像数据解码出大脑思考的内容？
- 能否借鉴大脑的认知机理改进现有的大语言模型？

➤ 100 – 47%
 ➤ 2500 – 97%+
 ➤ 4000 – 99%+

提 纲

1. 引言

 2. 数据集构建

3. 脑神经编解码

4. 结束语

2. 数据集构建

(1) 汉语词汇级的神经影像数据(Sci. Data, Nov.2022)



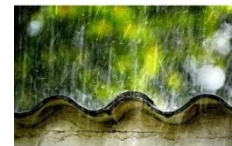
11位被试的母语均为汉语，每个词看3s，间隔2s(十字空屏)。每个人每次看2小时，672个词全部看完。

● 实验刺激

选取《同义词词林》中每个类别的代表性词汇，包括名词、动词、形容词，共计672个，10类，每个词汇呈现6遍。



下雨



下雨



下雨



下雨



下雨



下雨

下雨

上课

奔放

太阳

思考

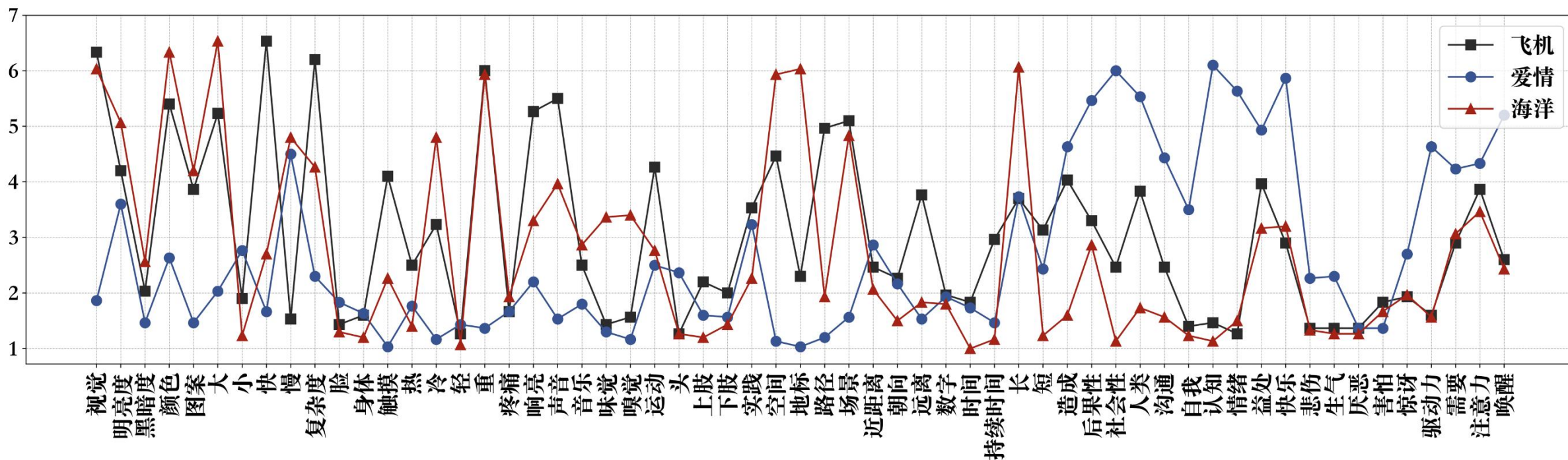
- 大脑每个体素编码了什么语义信息?
- 不同(类型)词汇的大脑表征有何异同?
- 词汇关系是如何在大脑中编码的?
- 大脑语义编码与语言学中概念表征的异同在哪里?



2. 数据集构建

● 语义特征标注

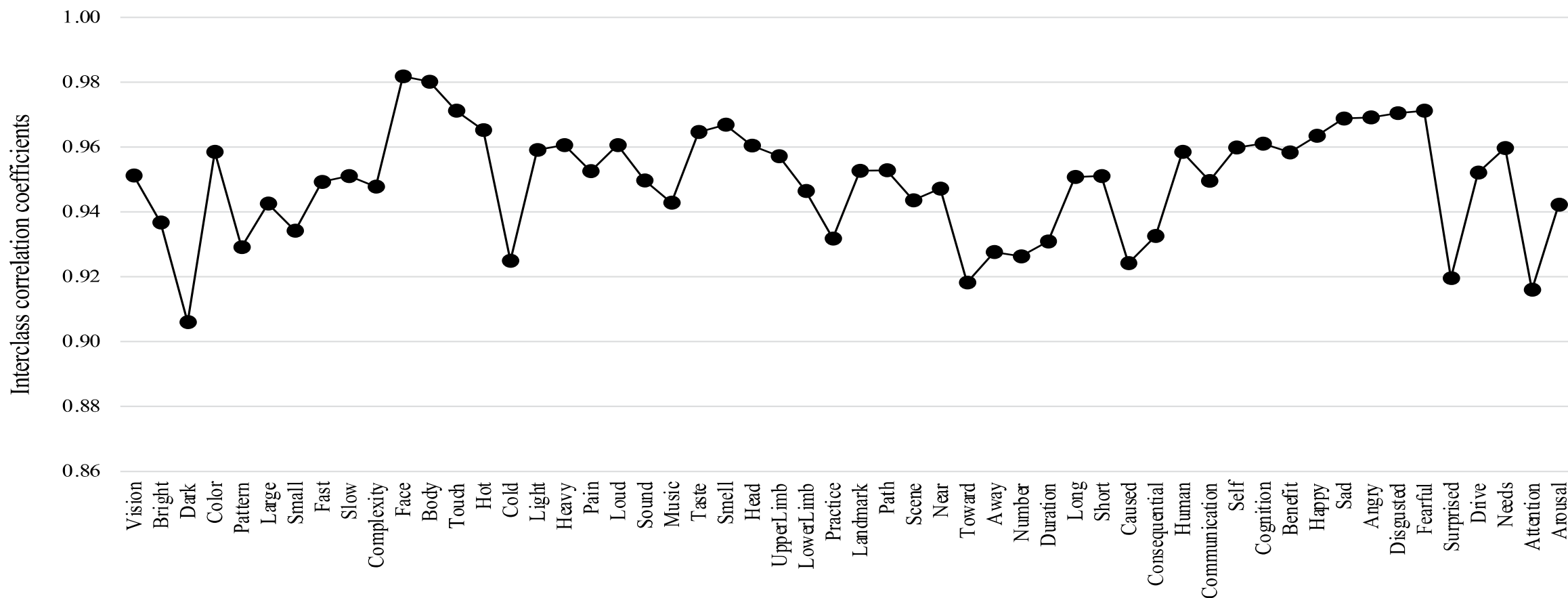
- ◇ 让30位汉语母语被试对这672个中文词汇进行标注；
- ◇ **54个语义维度**: 感知、空间、时间、社会性、情绪等。



从 感觉运动特征 到 抽象特征

2. 数据集构建

- 不同人对属性标注的分数一致性



ICC results of all semantic features.

2. 数据集构建

(2) 连续故事的神经影像数据(Sci. Data, Sept. 2022)

● 实验刺激

脑磁图仪



+ 磁共振
成像仪



- 60个中文故事，长度在600~900个词之间，总计40145个中文单词，词汇量为10152，每个词出现的频次范围为1-210；
- 音频摘自《人民日报评论》；
- 已采集12个被试的9个小时fMRI和6个小时MEG数据。

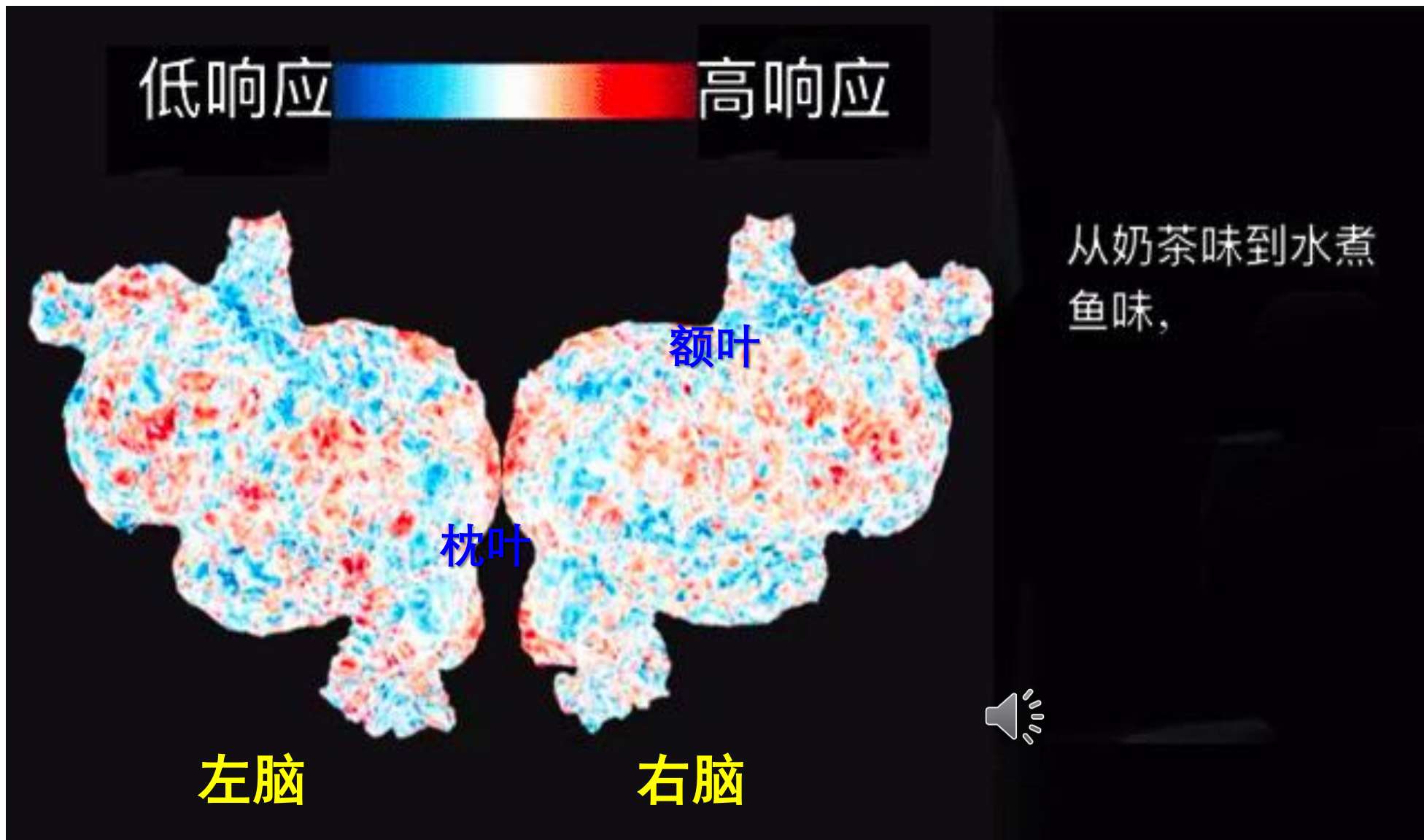


“从奶茶味到水煮鱼味，从折耳根味到钵钵鸡味，刷新味蕾的不仅只有月饼、酸奶，薯片也爆款频出，甚至一袋难求...”

2. 数据集构建

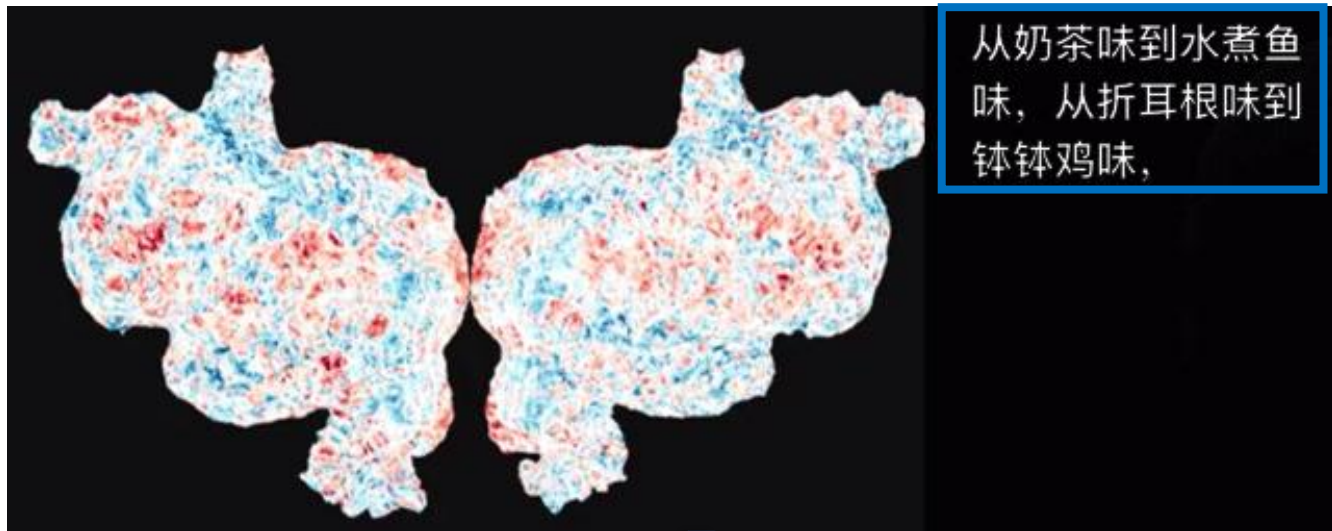
- 故事理解
响应

大脑皮层展
开后示意图



2. 数据集构建

● 分析



从/p 奶茶味/n 到/p 水煮/n 鱼/n 味/n 、 /wp 从/p 折耳根/n 味/n 到 /p 钵钵鸡/n 味/n 。 /wp



- 不同语言特征（词义、词法、句法、篇章关系等）是如何在大脑中编码的？
- 不同语言特征的时间响应关系是什么？
- 大脑故事表征过程与通用语言模型之间有什么联系？



2. 数据集构建

(3) 大规模汉语语义评分及其计算扩展数据集(Sci. Data, Feb. 2023)

- 被试: 30位汉语母语被试。
- 6个语义维度: 视觉表象性、动作表象性、空间性、时间性、社会性和情绪性。
- 主观标注数据集: **17940** 个中文词汇 (**12814 + 4915 + 211**)。

① 维基百科中文语料库

+ 中国语言数据联盟语料库(Chinese LDC)

} $\Rightarrow$ 三位独立评分员 **12814**

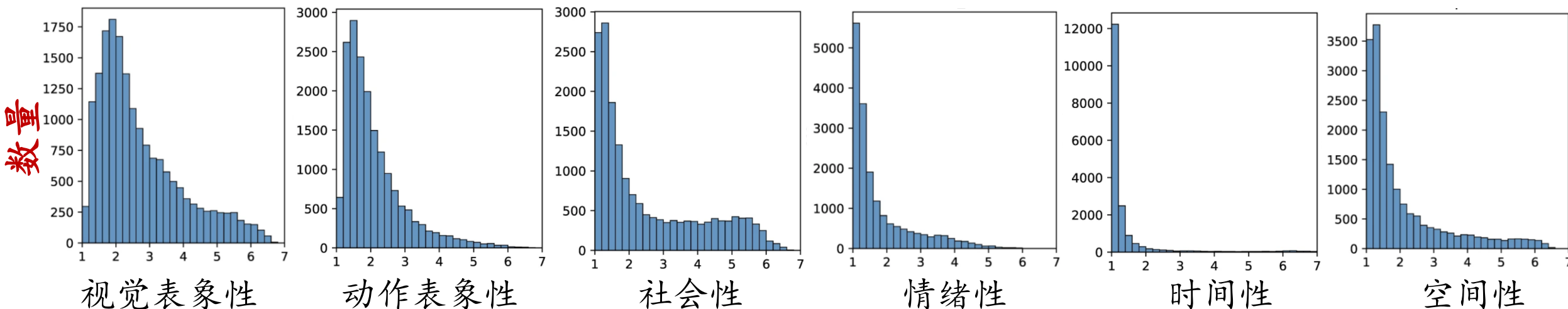
② 我们发布的两个 fMRI数据集(Sci. Data., 672) + 多模态故事 **4915**

③ 威斯康辛医学院和普林斯顿大学的英语语义评价实验词汇的中文翻译 **211**

2. 数据集构建

(3) 大规模汉语语义评分及其计算扩展数据集(Sci. Data, Feb. 2023)

- 被试: 30位汉语母语被试。
- 6个语义维度: **视觉表**象性、动作表象性、空间性、时间性、社会性和情绪性。
- 主观标注数据集: **17940** 个中文词汇 (**12814 + 4915 + 211**)。



- 通过计算方法对拓展数据集进行语义标注: **1427992**个中文词汇(新华网)和**1515633**个英文词汇(维基百科)



2. 数据集构建

(4) 多模态认知基准数据集(*Language Resource Evaluation, LRE*)

- 5个中文数据集和4个英文数据集: 词级别fMRI、语篇级别的fMRI和MEG; 词级别的语义标注数据; 句子级别的眼动数据。

用于评测LLMs 的认知能力

Language	Modality	Source	Stimuli	Subject	Unit	Tokens
Chinese	word semantic rating	CRSF	text	30	word	672
	word fMRI	CRSF	text	11	word	672
	eye-tracking	Zhang et al.	text	1718	sentence	170,331
	discourse fMRI	SMN4Lang	audio	12	discourse	52,269
	discourse MEG	SMN4Lang	audio	12	discourse	52,269
English	word semantic rating	Binder et al.	text	30	word	535
	word fMRI	Pereira et al.	text	15	word	180
	eye-tracking	ZuCo	text	30	sentence	36,767
	discourse fMRI	Zhang et al.	audio	19	discourse	47,356

2. 数据集构建

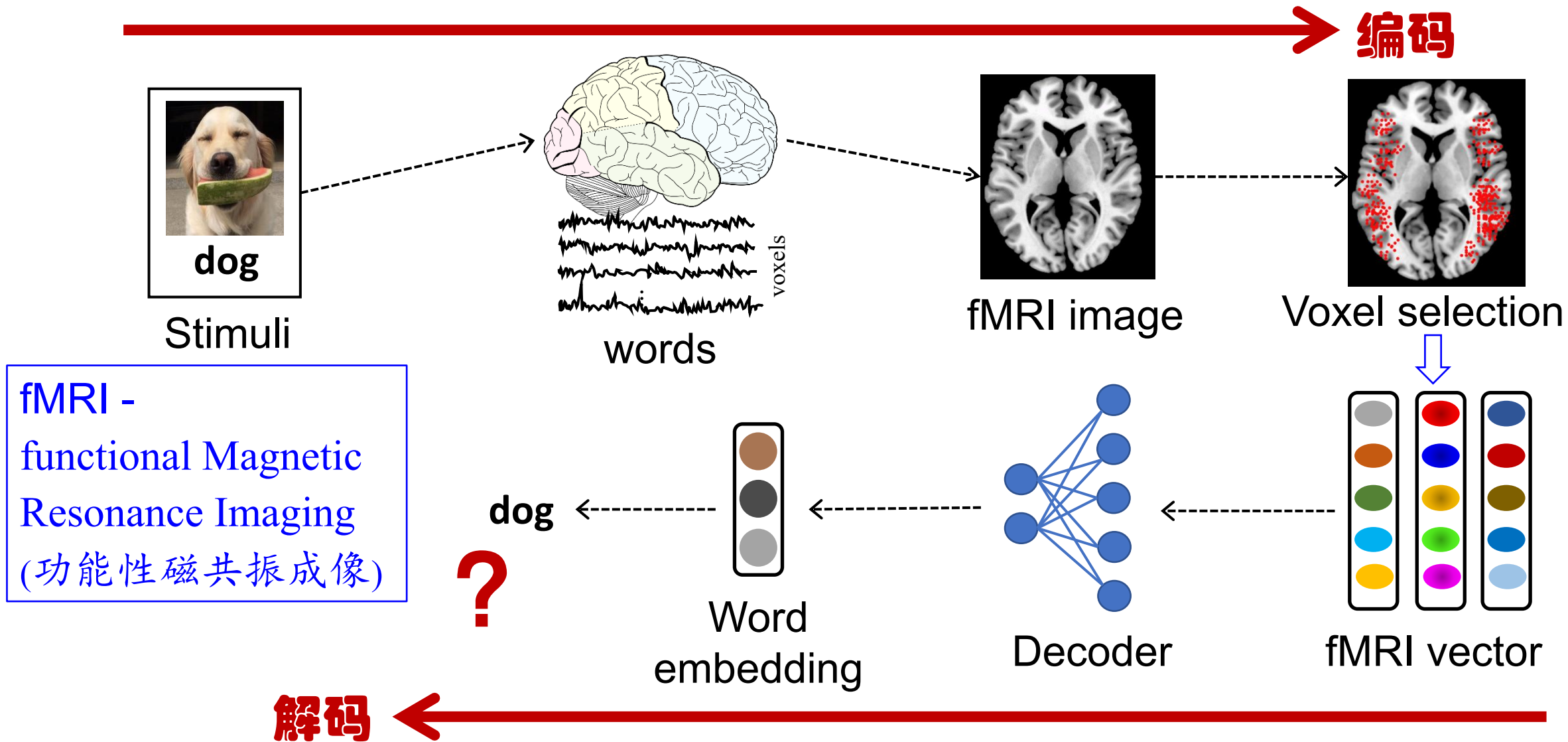
◆ 相关论文

- ① S. Wang et al., An fMRI Dataset for Concept Representation with Semantic Feature Annotations. *Scientific Data*, <https://doi.org/10.1038/s41597-022-01840-2>
- ② S. Wang et al., A synchronized multimodal neuroimaging dataset for studying brain language processing. *Scientific Data*, (2022) 9: 590. <https://doi.org/10.1038/s41597-022-01708-5>
- ③ S. Wang et al., A large dataset of semantic ratings and its computational extension. *Scientific Data*, <https://doi.org/10.1038/s41597-023-01995-6>
- ④ Y. Zhang et al. A Comprehensive Neural and Behavioral Task Taxonomy Method for Transfer Learning in NLP. In *Findings of IJCNLP-AAACL 2023*, pp. 233–241
- ⑤ Y. Zhang et al. MulCogBench: A Multi-modal Cognitive Benchmark Dataset for Evaluating Chinese and English Computational Language Models, *Language Resources and Evaluation (LRE)*, 1-24. <https://doi.org/10.1007/s10579-025-09843-2>

提 纲

1. 引言
2. 数据集构建
-  3. 脑神经编解码
4. 结束语

3. 脑神经编解码



3. 脑神经编解码

- ◆ 问题：如何更好地分析和预测大脑的神经激活？
 - 已有的研究使用的数据集要么规模太小，要么只包含来自单一刺激范式或单一语言单元的刺激；
 - 已有的研究大多数集中在印欧语言，特别是英语；
 - 现有的神经编码研究过度关注神经语言模型预测大脑激活的能力，忽视了基于心理学模型的编码能力；
 - 借助于大语言模型等AI技术能否对脑神经编码机制进行深入探索，构建出大脑的语义图谱？
 - 如何借助与AI技术实现基于影像和脑电数据的准确解码？

3. 脑神经编解码

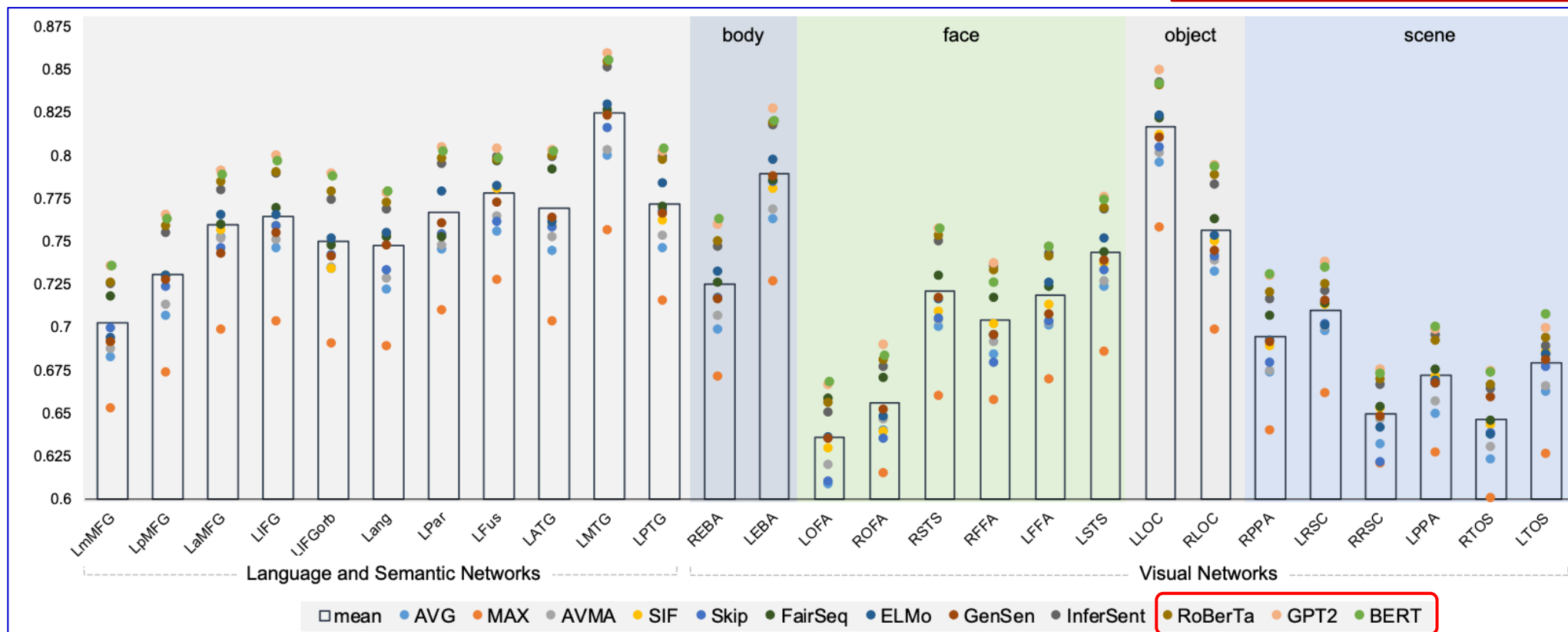
(1) 分布式文本表示

模型分类		名称	结构基础
非结构化模型	朴素池化	① 平均池化(AVG)	平均池化
		② 最大池化(MAX)	最大池化
		③ 联合池化(AVMA)	拼接平均池化和最大池化
	复杂池化	④ SIF	加权平均池化
结构化模型	无监督模型	⑤ Skip-thought	LSTM
		⑥ FairSeq	CNN
		⑦ ELMo	CNN-BiLSTM
		⑧ BERT	Transformer
		⑨ GPT2	
		⑩ RoBerTa	
	监督模型	⑪ GenSen	BiLSTM
		⑫ InferSent	CNN-BiLSTM

3. 脑神经编解码

● 分布式表示模型在脑语言及视觉网络上的编码准确率

IEEE TNNLS, No. 2, 2020



- 候选模型在大脑左颞叶中回(LMTG)和左枕叶复合体(LLOC)上编码表现突出；
- 基于Transformer 的模型在语言网络上的编码准确率显著优于其他模型；
- 应用最广泛的平均方法，其编码表现仅排在所有候选模型的中游。



3. 脑神经编解码

(2) 语言模型 vs. 心理学模型

模型分类		名称	维度	层数	结构基础
基于心理学的模型		① 局部统计模型 (LSM)	300	\	简单共现
		② 网络拓扑模型 (NTM)	256	\	拓扑关系
		③ 具身理论模型 (EBM)	6	\	语义评分
学习驱动的语言模型	上下文感知模型	④ BERT _{base}	768	12	Transformer
		⑤ GPT2 _{ori}	768	12	
		⑥ BERT _{large}	1024	24	
		⑦ GPT2 _{med}	1024	24	
	上下文无关模型	⑧ Word2Vec	300	1	神经网络
		⑨ GloVe	300	\	矩阵分解

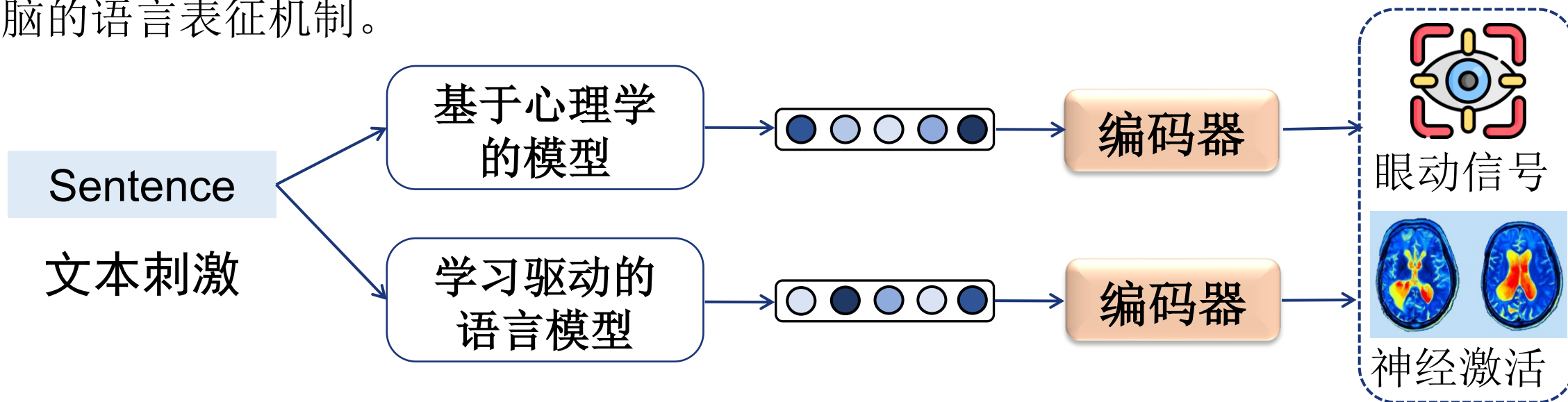
3. 脑神经编解码

➤ 数据集

— 多模态认知基准数据集(LRE)

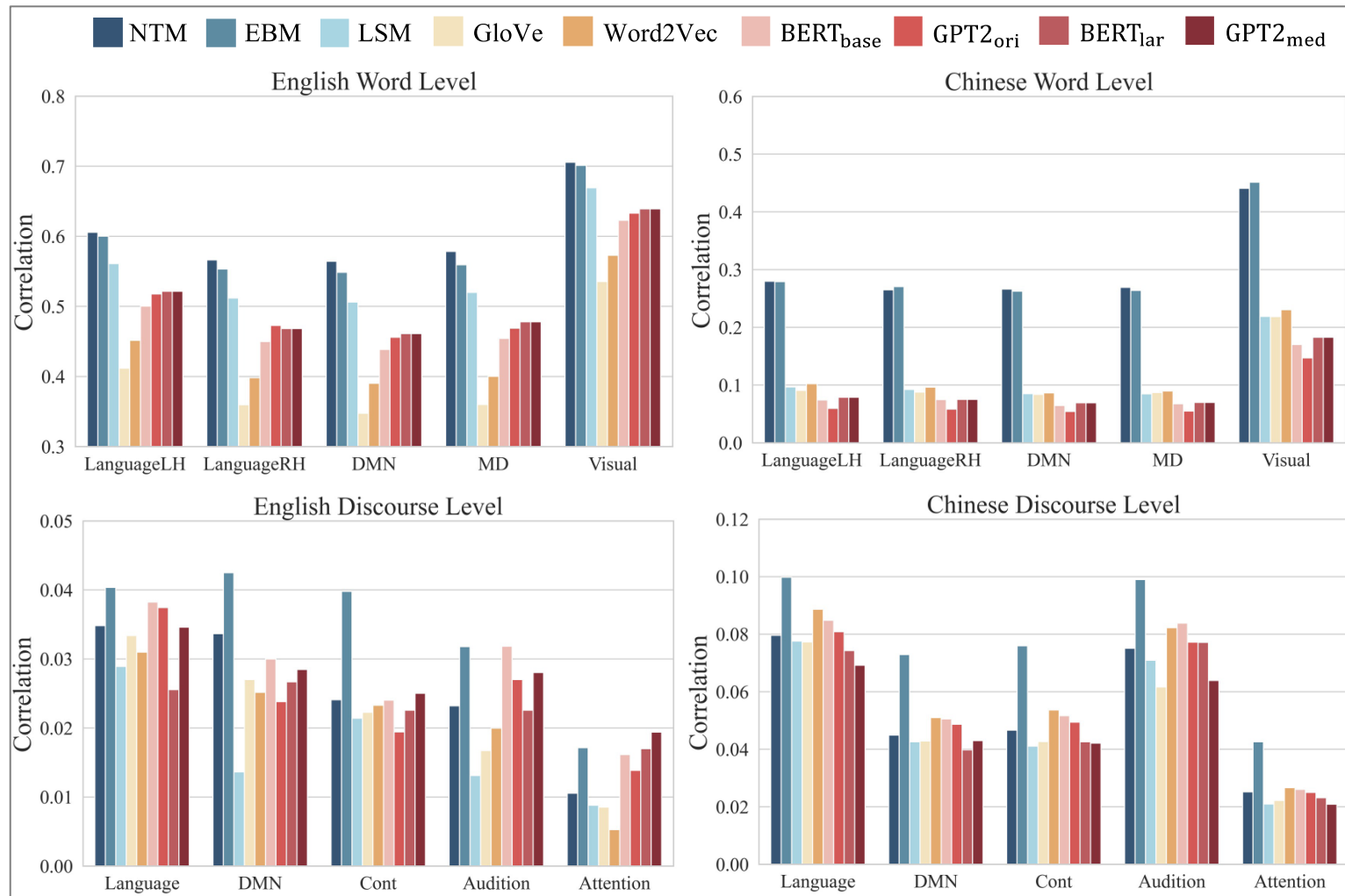
➤ 神经编码方法

目标: 通过语言刺激的变化来预测对应各脑区、体素点神经活动, 进而帮助解析人脑的语言表征机制。



3. 脑神经编解码

➤ 结果-1：不同类型模型在5个主要脑网络上的编码准确率



◇在词级别的fMRI上，**网络拓扑模型(NTM)**和**具身理论模型(EBM)**编码表现突出；

◇在语篇级别的fMRI上，**具身理论模型(EBM)**获得了相对较高的编码准确率。

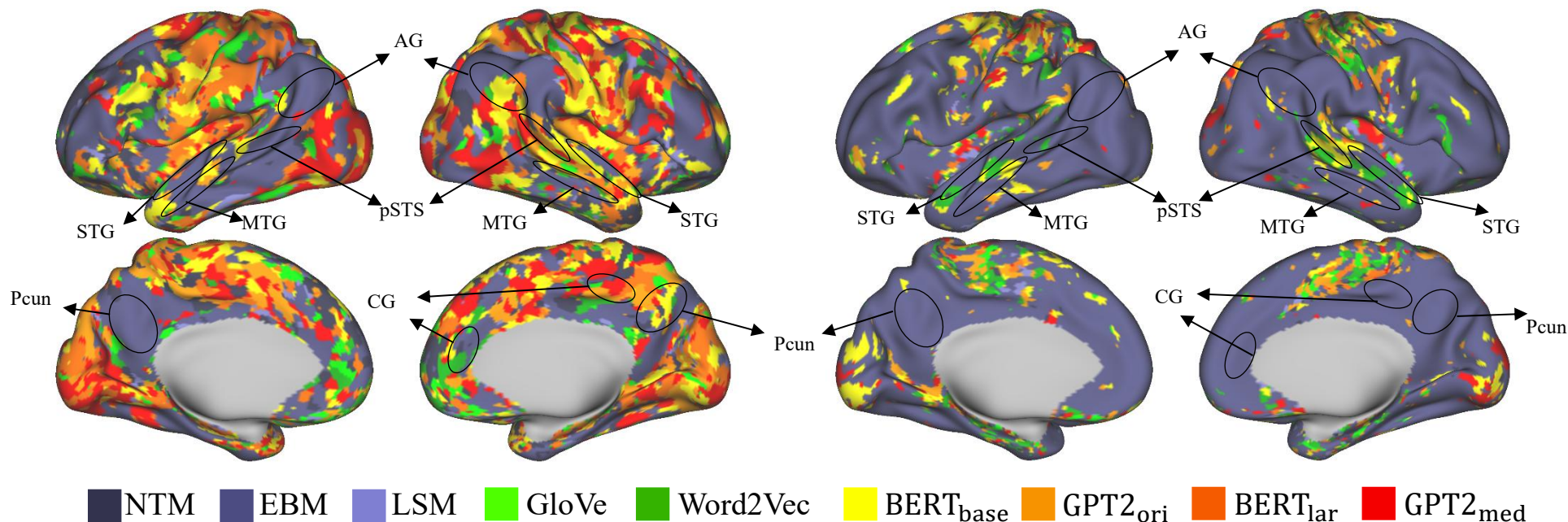
具身理论模型与大脑语义理解机制相关性更强/ 大脑语义理解机制相一致？

3. 脑神经编解码

➤ 结果-2：不同语言模型在全脑体素级别上的编码表现

◇ 每个体素用能够最好预测它的语言模型进行标记

Proc. of Cog.Sci. Aug. 2024.



- ◇ 具身理论模型 (EBM) 在大多数与语义处理相关的ROI中实现了优越的编码性能；
- ◇ GPT2_{med} 在外侧枕叶皮质 (LocG), 岛回 (INS) 和顶上小叶 (SPL) 上编码表现突出；
- ◇ BERT_{large} 捕获了角回 (AG), 额中回 (MFG) 和颞上后沟 (pSTS) 上的神经激活。

3. 脑神经编解码

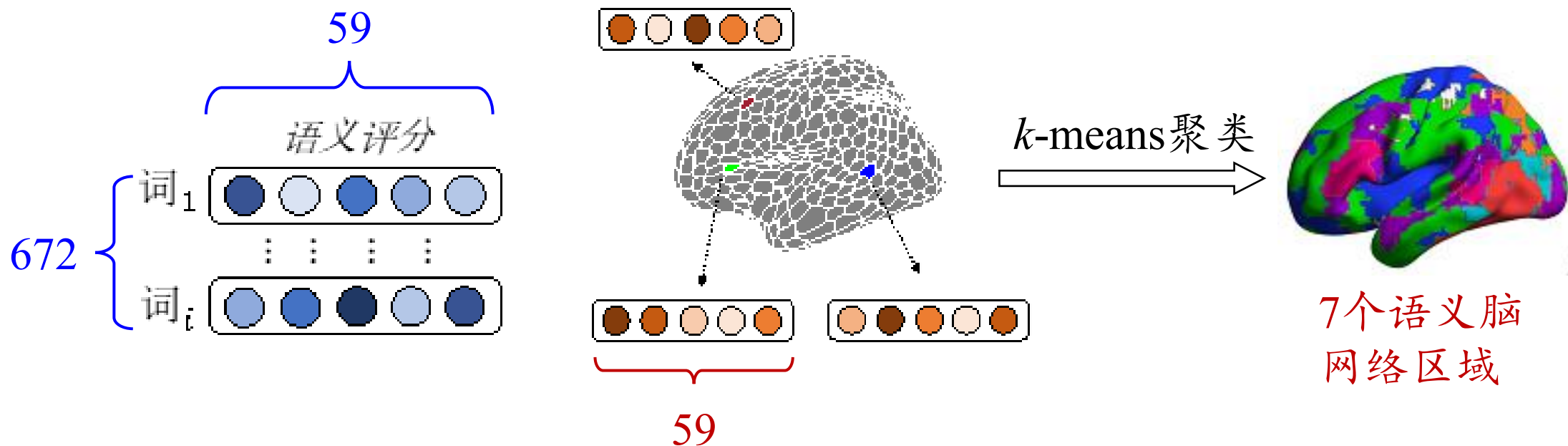
(3) 基于fMRI数据构建语义表征功能驱动的大脑皮层网络分区—语义图谱

- 语言理解涉及语义表征与多种认知功能的协同，其神经活动反映了多种认知功能，构建语义网络需从神经活动中分离出语义相关信息。
- 对语义网络进行了可靠性与有效性测试，评估其稳定性与可扩展性。
- 探究每个语义网络的认知功能：
 - 考察多维语义表征功能；
 - 追踪语义信息的获取来源；
 - 分析语义网络在通用认知功能中的参与情况。

3. 脑神经编解码

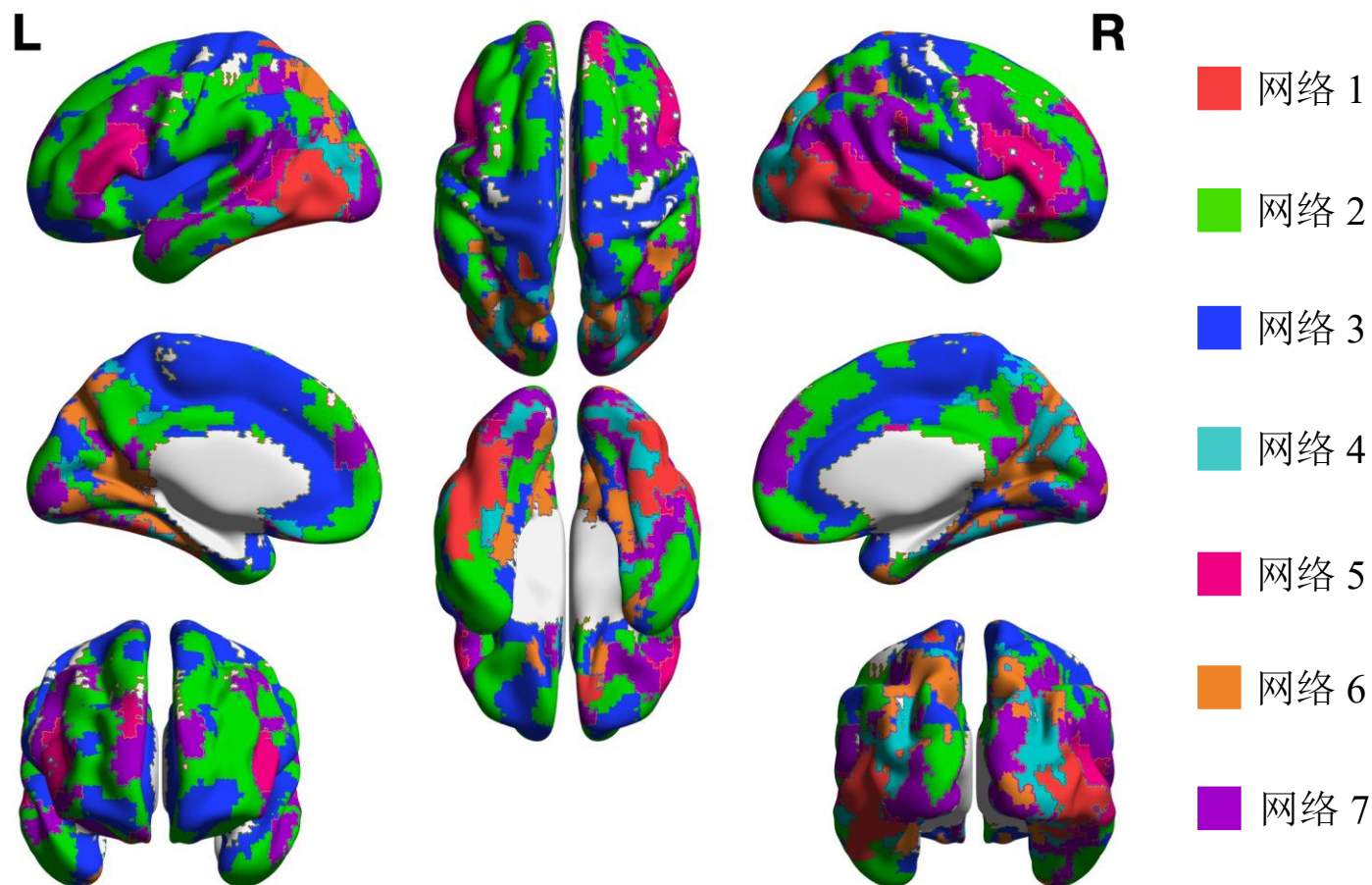
● 提出了一种基于特定认知功能的大规模脑网络划分方法

- **目的:** 通过表征相似性编码将大脑每个区域映射到可解释语义维度模型 (具身理论模型), 将每个脑区的语义信息分离出来, 生成语义表征矩阵, 并对该矩阵进行聚类得到大脑语义表征网络分区。



3. 脑神经编解码

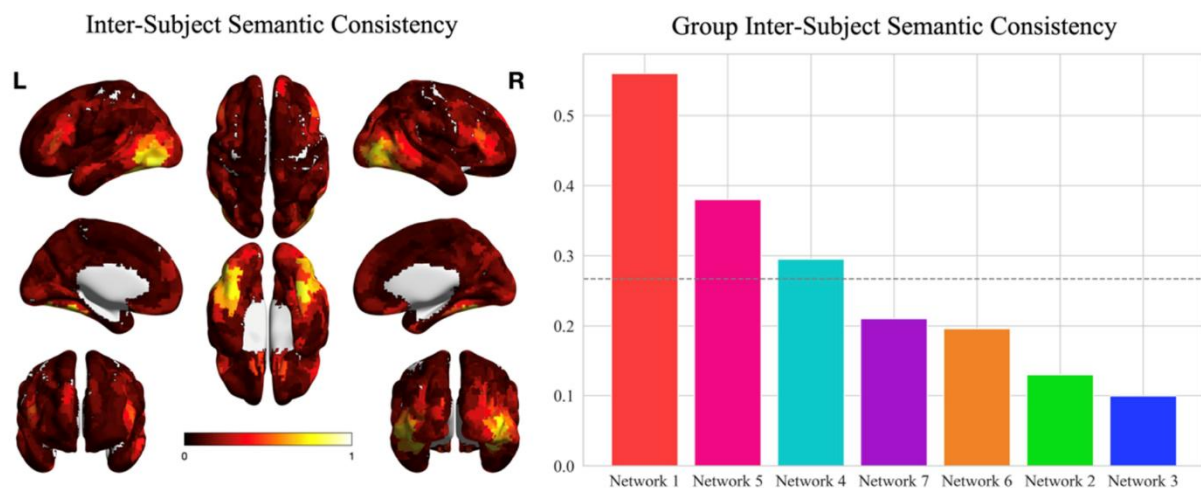
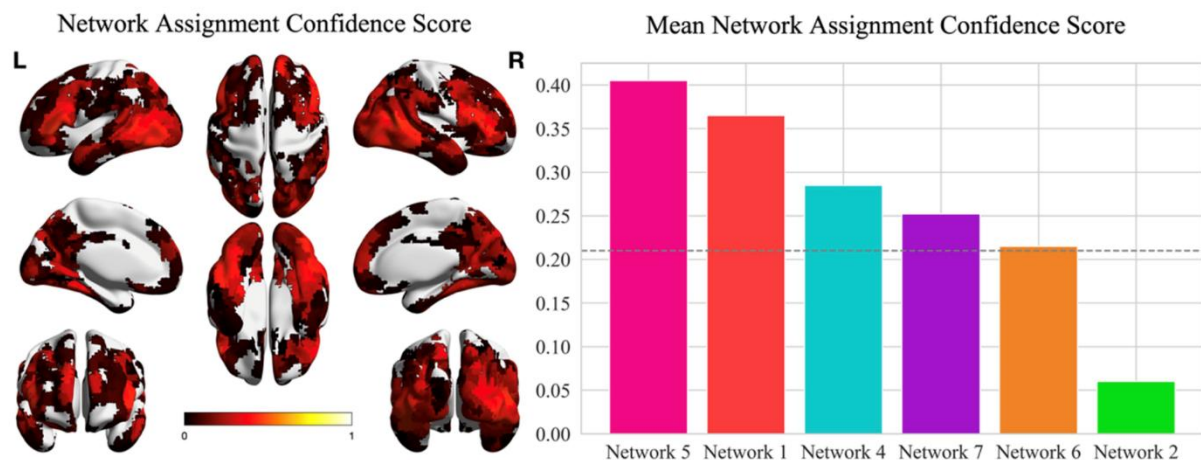
➤ 大脑语义网络分区



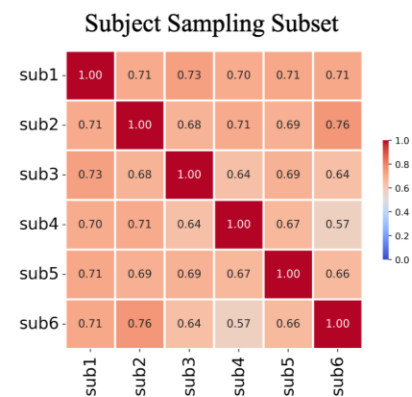
- ✧ 网络1和4位于颞叶和枕叶的腹侧与背侧区域，涉及**物体语义**；
- ✧ 网络6涵盖内侧颞叶及顶枕皮质，参与**空间语义**编码；
- ✧ 网络7与社会相关脑区高度重叠 (背外侧前额叶皮质、颞顶交界处、前颞上沟和后扣带皮层)，参与**社会语义**表征；
- ✧ 网络3中的**运动区**和**感觉运动信息**相关。

3. 脑神经编解码

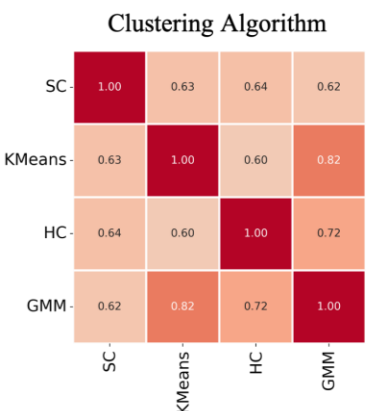
➤ 大脑语义网络的稳定性



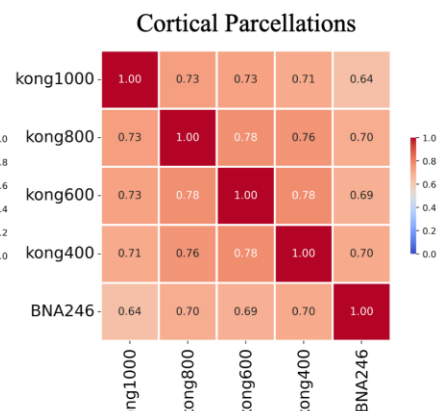
不同的被试



不同聚类算法



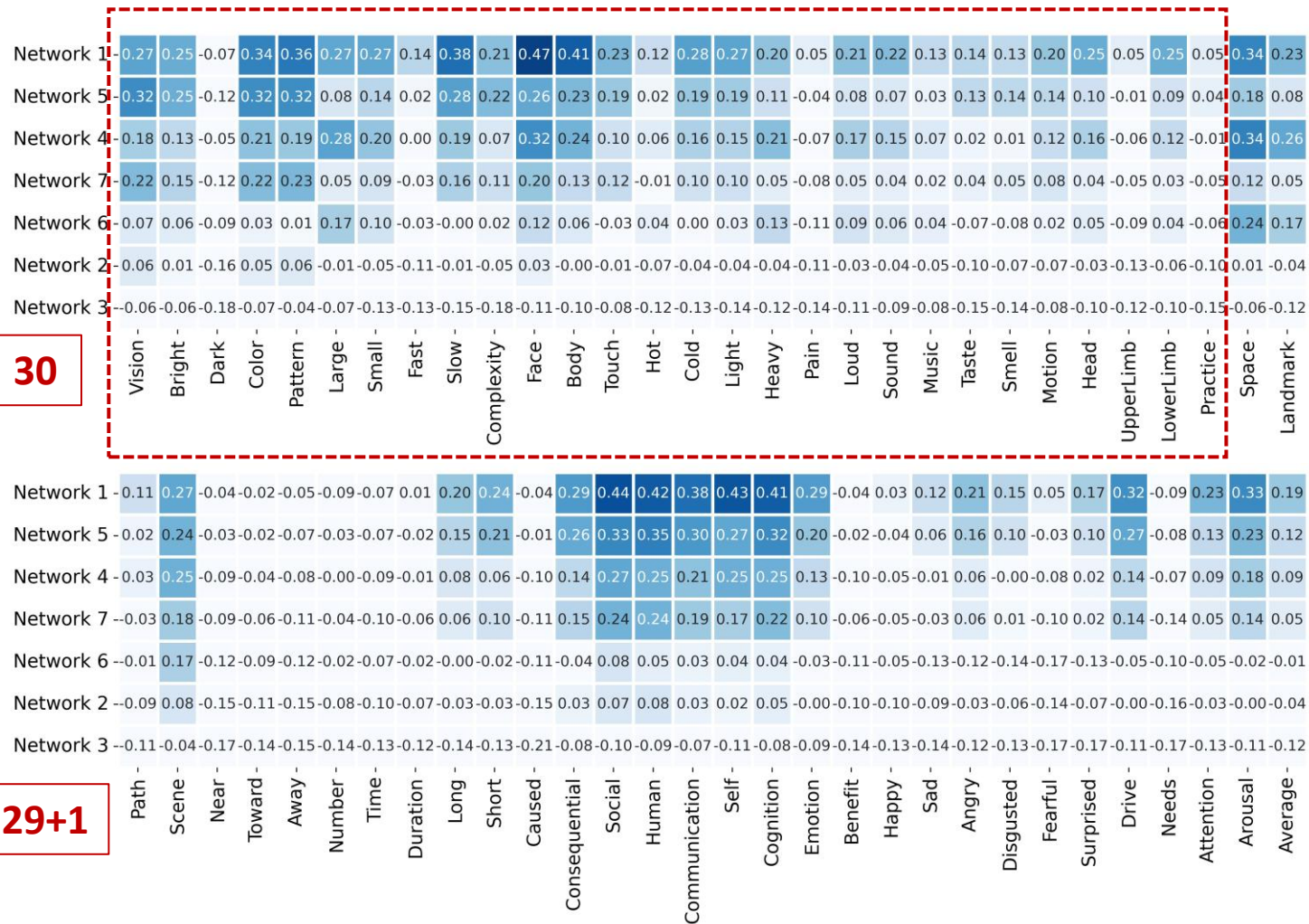
不同的皮层分区



- 使用不同被试子集、聚类算法以及皮层模版得到的语义网络间有极高的相似度；
- 网络置信度分数和主体间语义一致性均表明得到的语义网络有较高的稳定性。

3. 脑神经编解码

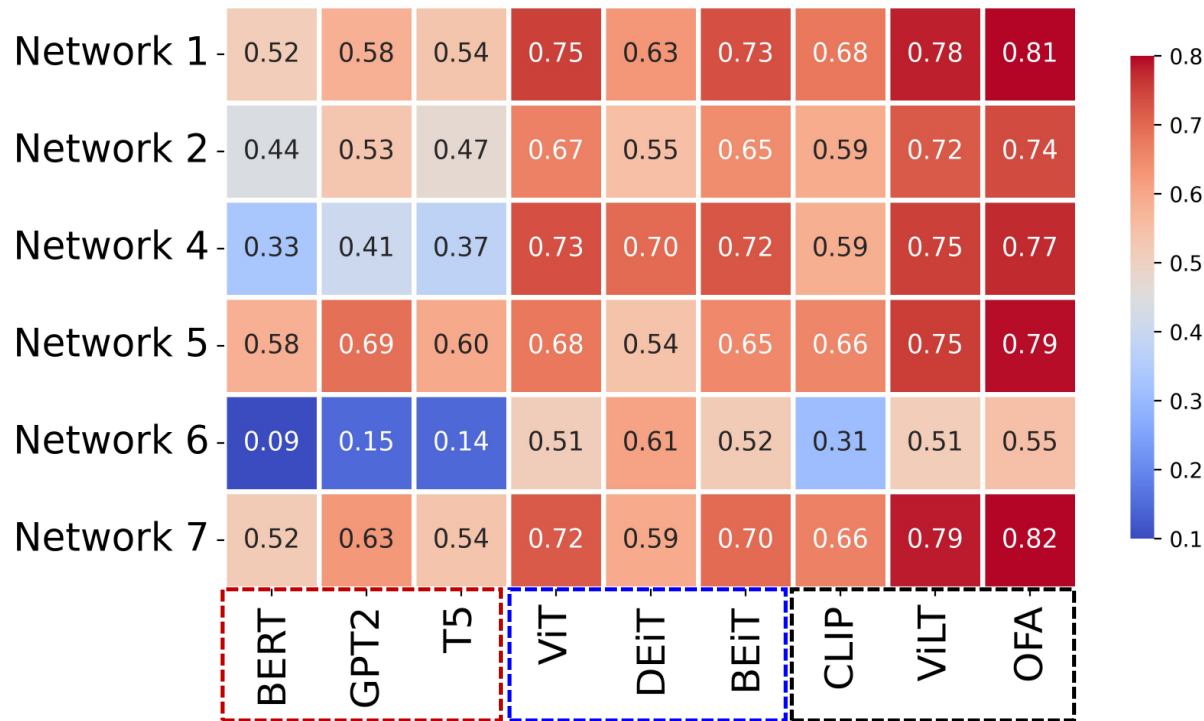
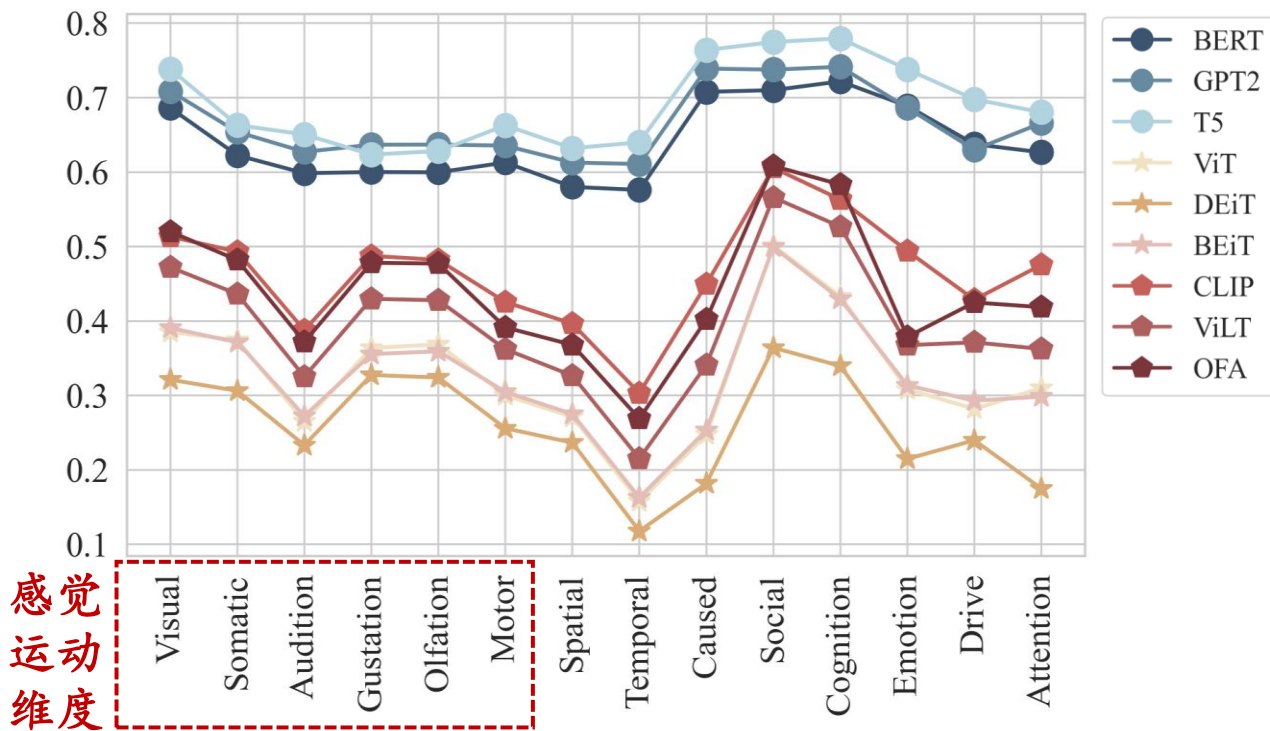
➤ 大脑语义网络的多维语义表征功能



- 在**感觉运动维度**方面，感觉信息在大脑中的编码程度高于其他运动信息；
- 非感觉运动维度**方面，大脑编码相对较多的是与社会和自我相关的语义信息。

3. 脑神经编解码

➤ 大脑语义网络语义信息的获取来源



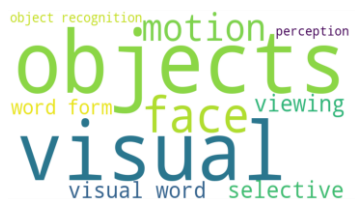
- 不同模态(语言/视觉/多模态)的模型在语义信息的分布上呈现出差异性;
- 网络 1、网络2、网络5和网络7通过**语言**和**视觉**模态获得语义知识，而网络4和网络6主要依赖**视觉**模态获取语义知识。

3. 脑神经编解码

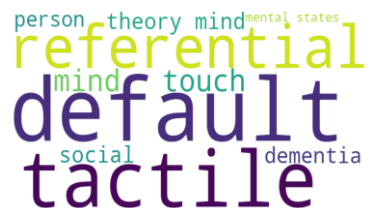
➤ 大脑语义网络在通用认知功能中的参与情况

- 使用Neurosynth对每个语义网络进行解码分析

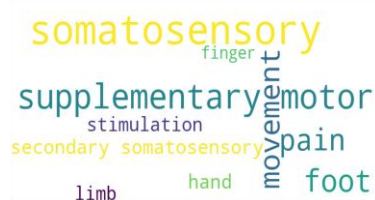
网络 1



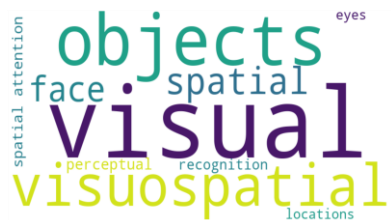
网络 2



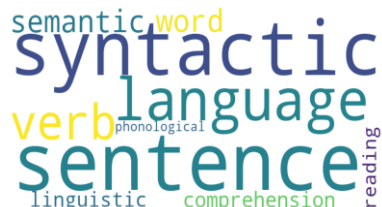
网络 3



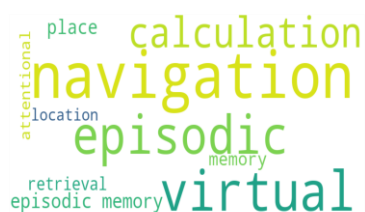
网络 4



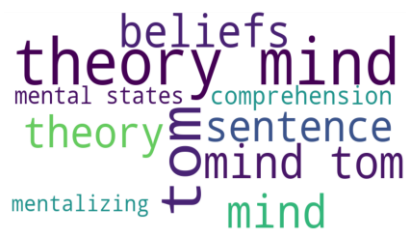
网络 5



网络 6



网络 7



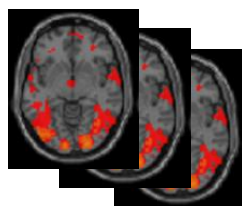
- 网络1、4，专注于视觉处理，主要识别人脸、物体和其他视觉刺激；
- 网络2，社会认知网络，涉及高阶认知功能，包括默认模式处理、触觉感知等；
- 网络3主要与躯体感觉和运动功能相关；
- 网络5，语言处理网络，涉及语言处理，包括语法、语义和音系；
- 网络6参与导航、情景记忆和空间处理。

NeuroImage, Feb. 2025

3. 脑神经编解码

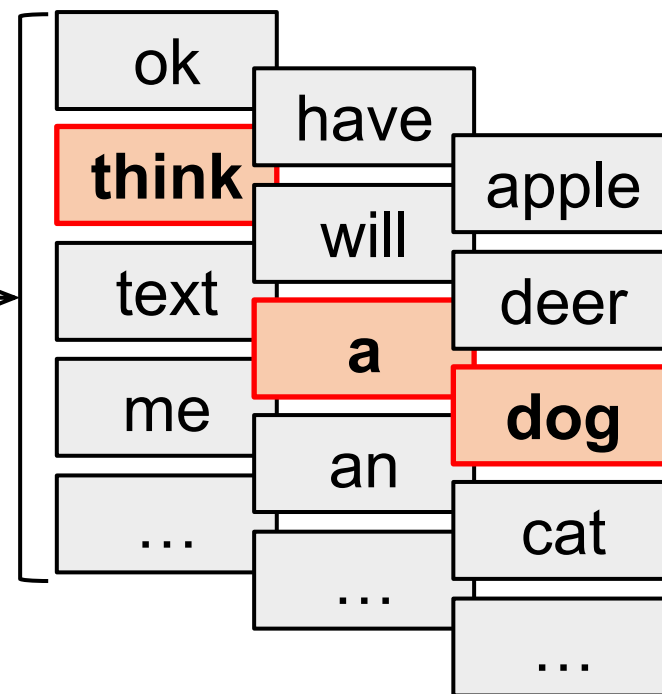
(4) 句子级解码：借助预训练语言模型实现基于fMRI数据的句子生成

fMRI2text Task



Decoder

Input: continuous fMRI image



Output: target sentence

Proc. of NAACL 2024

3. 脑神经编解码

◆ 相关论文

- ① X. Zhang et al. Is the Brain Mechanism for Hierarchical Structure Building Universal Across Languages? An fMRI Study of Chinese and English. *Proc. of EMNLP 2022*
- ② X. Zhang et al. How Does the Experimental Setting Affect the Conclusions of Neural Encoding Models? *Proc. of LREC 2022*, pp. 6397–6404
- ③ X. Zhang et al. Probing Word Syntactic Representations in the Brain by a Feature Elimination Method. *Proc. of AAAI 2022*
- ④ J. Sun et al. Neural Encoding and Decoding with Distributed Sentence Representations, *IEEE TNNLS*, Vol.32, No.2, pp.589-603, 2020
- ⑤ J. Sun et al. Distill and Replay for Continual Language Learning. *Proc. of COLING 2020*, pp. 3569-3579
- ⑥ Y. Zhang et al. Navigating Brain Language Representations: A Comparative Analysis of Neural Language Models and Psychologically Plausible Models. In *Proc. of the Annual Meeting of the Cognitive Science Society*. Vol. 46. 2023

3. 脑神经编解码

- ⑦ Y. Zhang et al. A simple clustering approach to map the human brain's cortical semantic network organization during task. *NeuroImage*, Feb. 2025
<https://doi.org/10.1016/j.neuroimage.2025.121096>
- ⑦ S. Zou et al., Cross-Modal Cloze Task: A New Task to Brain-to-Word Decoding. In *Findings of the Association for Computational Linguistics (ACL 2022)*, pages 648 - 657, May 22-27, 2022
- ⑧ S. Zou et al., Towards Brain-to-Text Generation: Neural Decoding with Pre-trained Encoder-Decoder Models. *Proc. of NeurIPS 2021 AI for Science Workshop*. December 6 - 14, 2021(online)
- ⑨ X. Zhao et al., MapGuide: A Simple yet Effective Method to Reconstruct Continuous Language from Brain Activities. *Proc. of NAACL 2024*

提 纲

1. 引言
2. 数据集构建
3. 脑神经编解码
-  4. 结束语

4. 结束语

- 大语言模型为探索人脑理解语言的神经机制提供了新的研究思路和技术手段，但是，不应因为追求新技术手段而放弃已有的经典方法；
- 借助于预训练语言模型根据 fMRI 数据直接生成词语，甚至在给定上下文的情况下生成连贯的自然语言文本句子，完全是可能的，而且有令人期待的结果；
- 借助于对大脑功能和认知机理的认识，改进大模型结构，压缩大语言模型的规模，提升大模型在低资源语言上的处理能力，甚至改进模型结构等，都是值得探索的技术路线。

谢谢!
Thanks!

