

认知神经科学

语义编码与多模态对齐：综述与实验报告

基于故事听觉 fMRI 的预训练模型特征比较与相关工作整合

Semantic encoding and multimodal alignment: a review and experimental report

报告人 潘宇轩

学 号 2023K8009991004

院 系 人工智能学院

专 业 人工智能

日 期 2026 年 1 月 23 日

摘要

本报告将“相关工作综述”与“当前项目实验结果”合并呈现。综述部分围绕自然语言刺激下的语义脑图绘制 [4, 12]、预测编码与集成建模框架 [9, 1, 5]、自监督语音表征与大脑对齐 [2, 6]、组合语义与上下文表征 [7, 3, 8, 11] 以及连续语义重构的解码方向 [10]，建立本实验的理论与方法背景。实验部分基于当前项目目录中已生成的结果文件撰写，所有数值均来自 `results/summary.csv`、`results/roi.csv` 与各模型目录下保存的相关图（`corr_layer*.npy` 以及融合的 `corr_*.npy`）。研究目标是在故事听觉 fMRI 数据上，对比文本模型、音频模型、多模态模型在不同层与不同时间窗口（TR 窗口）条件下的脑预测性能，并结合 ROI 统计与可视化图像分析区域偏好。实验采用统一的对齐、降维与时延展开流程，并在多被试上进行单次训练/测试划分的岭回归评估，从而得到可复现的全脑相关图。报告呈现两类图像：其一为统计图（模型对比与窗口趋势、ROI Top20），由 `report/scripts/make_figures.py` 读取现有结果自动生成；其二为脑图，可视化来自 `src/run_plot_corr_maps.py` 对保存的 corr map 进行映射绘制。

目录

1 第一部分：相关工作综述与研究动机	5
2 引言	5
3 预测编码理论及其争议	7
3.1 预测编码的基本思想	7
3.2 对预测编码证据的质疑	9
3.3 集成建模对预测性的支持	10
4 语义系统的脑映射与分布式表征	12
4.1 数据驱动绘制语义地图	12
4.2 语义类别与关系的分布式编码	13
4.3 地图与网络的意义	15
5 学习机制与模型：自监督语音与 Transformer	15
5.1 自监督语音模型 wav2vec 2.0.	15
5.2 Transformer 的计算与功能专化	16
6 组合语义与 supra-word 表征	18
7 逆向工程：从神经信号重构连续语义	19
8 方法论：编码模型、评价指标与统计检验	20
8.1 编码模型：岭回归与分带岭回归	20
8.2 噪声天花板与显著性检验	20
9 跨语言与跨模态比较	21

10 讨论：综合视角与未来方向	22
10.1 预测还是特征学习？	22
10.2 分布式语义系统与网络化视角	23
10.3 模型特性与大脑对应	23
10.4 组合语义的新挑战	23
10.5 跨文化与多语言的扩展	23
11 第二部分：实验设计、方法与结果	25
12 背景、目标与本报告范围	25
13 数据、对齐表与 TR 级刺激构建	26
14 预训练模型、层选择策略与特征提取实现	27
14.0.1 模型集合与本次报告覆盖范围	27
14.0.2 层选择：按相对深度等比例取样	28
14.0.3 文本特征：上下文窗口与双层池化	28
14.0.4 音频特征：TR 窗口切分与帧级池化	29
14.0.5 多模态特征：模型内部融合与 TR 对齐	29
15 编码模型、评价指标与输出结构	29
15.0.1 从 TR 特征到 BOLD：线性编码模型的形式化	29
15.0.2 BOLD 延迟建模：PCA 降维与 FIR 延迟展开	30
15.0.3 数据划分与多被试汇总	30
15.0.4 评价指标与 corr map 的含义	31
15.0.5 输出文件结构与可追溯性	31
16 文本模型结果：层次对齐与区域分布	32
17 音频模型结果：TR 窗口效应与最佳层比较	35
18 多模态模型结果：Whisper	35

目录

认知神经科学

19	文本 + 音频融合结果：覆盖范围、最优配置与层交互结构	43
20	ROI 分析与综合讨论	44
21	结论、局限与后续可扩展方向	47
A	附录：结果文件位置与复现说明	49

1

第一部分：相关工作综述与研究动机

本部分用于建立实验的研究语境与方法参照。自然叙事刺激下的语义地图研究表明，语义信息在皮层上呈现分布式组织，并跨越传统语言区与默认模式网络等大尺度系统 [4, 12]。在模型侧，集成对齐研究将不同语言模型与脑数据进行系统比较，指出架构、训练目标与层级表示共同影响对齐程度 [9, 1, 5]。在语音与多模态方向，自监督语音表征能够在较少标注下形成与人类加工相近的表征梯度 [2, 6]。在语义组成与上下文依赖方面，ELMo、BERT、GPT-2 等模型提供了不同的上下文表示机制，为探查“组合意义”在脑中的可测表现提供了可操作的特征空间 [7, 3, 8, 11]。与此同时，连续语义重构工作展示了以语义而非字面文本为目标的解码可能性，为“编码模型”与“解码模型”的统一评估提供了参考 [10]。

2

引言

近年来，随着深度学习技术的突破，人工神经网络在自然语言处理（NLP）领域获得了前所未有的成功。与此同时，认知神经科学的发展使得我们可以通过功能磁共振成像（fMRI）、电生理记录（ECOG）或脑磁图（MEG）等技术测量大脑在理解语言时的动态响应。将这两条研究路径结合起来，比较神经网络模型的内部表征与大脑活动，既有助于解释模型为何有效，也有助于揭示人脑处理语言的规律。特别是语言模型的训练目标往往是预测下一个词，这与某些理论认为大脑通过预测未来输入来减少处理负荷的观点不谋而合。然而，模型解释大脑的成功是否意味着大脑真的在实施同样的预测机制，仍存争议。

另一方面，数据驱动的语义地图绘制为理解语义系统提供了全新的视角。在自然语言刺激下，大脑的语义区域如何分布？语义类别和语义关系如何在皮层中表示？这些问题的解答不仅深化我们对语言认知的理解，也可能反过来指导人工模型的改进。本综述不按单篇文章分别介绍，而是围绕核心主题整合现有研究成果：首先概述预测编码理论及其在语言领域的支持与质疑；随后总结利用神经语言模型建立大脑编码模型的集成建

模工作；接着讨论语义系统的脑映射，包括语义类别与语义关系的分布式表征；之后阐述不同学习机制下模型与大脑的对齐，如自监督语音模型和 Transformer 的注意头分析；进一步介绍组合语义（supra-word）表征的发现；最后探讨跨语言、跨模态比较的意义，并在讨论部分提出未来研究方向。通过这种综合视角，我们旨在呈现大脑语言处理与人工神经模型互动的完整景观，并指出当前研究的关键问题和潜在路径。

本领域的交叉研究不仅推动认知科学的发展，也对人工智能的设计产生深远影响。一方面，神经网络模型已从仅追求任务性能的“黑箱”演进为可以解释人类数据的认知模型。通过将脑成像数据纳入模型评测，研究者得以揭示模型所学的语言表示是否具有生物学真实性，从而指导模型架构和训练目标的调整。另一方面，大脑科学家利用模型生成具备特定语言属性的刺激，设计实验以验证大脑的语义组织、加工顺序和层次结构。例如，模型可以产生具有不同预测难度或不同语法依存结构的句子，用于探索大脑在预期违背或结构复杂度下的反应模式。正如后文所述，这种双向互动正在形成新的研究范式，促使人工智能和神经科学共同迈向更高水平的理解。

近年来，随着脑成像技术的进步，研究者不再局限于通过行为数据推断认知过程，而是能够在毫秒乃至更高时间分辨率上观察大脑对语言刺激的响应模式。结合深度学习模型，这种方法提供了一个跨尺度的桥梁，使我们能同时访问神经元群体的活动和算法级别的信息。与此同时，人工神经网络的内部表征从简单的线性结构发展为复杂的层次化系统，这些系统在训练过程中自发产生语法树状结构、意象表征乃至对世界的统计认识。对比这些表征与脑成像数据，可以揭示哪些结构是普遍存在于人类语言认知中的通用属性，哪些是模型特定的产物。

此外，围绕人工智能的伦理与实际应用问题也提醒我们，理解模型如何与人类认知对齐具有重要意义。语言模型被广泛用于教育、医疗和政策决策等领域，若忽略其与人类理解方式的差异，可能导致误解或偏差，甚至出现严重的社会后果。通过研究神经模型与大脑的对应，我们可以识别模型在语义推理、情感理解或常识推理方面的缺陷，并制定改进策略，例如在训练数据中引入多样化的语境、增加现实世界知识或限制模型的信任范围。总之，人工智能与神经科学的交叉研究不仅拓展科学边界，也对社会和伦理产生深远影响。

值得指出的是，语言不仅是词序列的累积，还包括丰富的形态学和语用学信息。许多语言使用屈折词缀、变化规则和音调来表达语法关系和时态、体、态等语义特征。例如，阿拉伯语的三根辅音根通过元音变换表达不同的词义，因纽特语的多重黏着导致单个词即可表示复杂句子。这些形态学特性为预测提供了额外线索：在富形态语言中，词干和词缀的组合给出句法框架，大脑可能利用形态学规则来缩小下一词的候选范围。未来的神经模型应考虑形态学多样性，不能仅依赖英语语料，否则在解析屈折或黏着语言时可能

无法对齐大脑处理。随着研究扩展到更多语言，比较不同形态类型的大脑活动将揭示语言特性对语义系统塑造的影响。

此外，语言的句法结构也对预测机制和语义分布有显著影响。语序固定的语言（如英语）与语序自由的语言（如拉丁语或俄语）在理解过程中的策略不同：在语序固定的语言中，听者和读者可以依赖固定的位置来识别主语、宾语和谓语，而在语序自由的语言中，需要依靠格标记和语义角色进行解析，这可能增加理解负荷并影响预测的层级。此外，汉语等话题优先语言常通过主题—述题结构组织信息，大脑必须在句子前段确定主题并预测后续信息。在某些美洲原住民语言中，主谓宾顺序可以灵活改变以突出新信息或对比，这种突出与焦点结构需要更复杂的语用推理。语言模型和脑编码研究应进一步探索语序变异如何影响预测和语义系统的组织，通过控制句法自由度和引入格标记等操作，比较不同语言的神经反应模式。

3

预测编码理论及其争议

3.1 预测编码的基本思想

预测编码（predictive coding）是一种流行的认知理论，它认为大脑通过不断生成对未来输入的预测来高效加工感知信息。大脑产生内部模型预测即将到来的刺激，并通过比较预测与实际输入的偏差（预测误差）更新内部模型，从而在多层次上实现感知和理解。这一框架最早在视觉系统提出，随后扩展到听觉和语言领域。语言过程中，大脑可能利用上下文信息预测接下来要出现的词语、语法结构或语义内容，减少加工负担。行为实验和脑电研究（如 N400、P600）提供了一些间接证据，表明当语句违背语法或语义预期时会出现特异的神经反应。

人工神经网络语言模型（NNLM）因其训练目标与预测未来词语相关，被视为测试预测编码假说的理想工具。语言模型通常通过最大化下一个词的概率来进行自监督学习，这使模型的隐藏表示内化了大量语法、语义和篇章结构信息。如果人脑确实在理解语言时进行类似预测，那么这些模型的表现应与人脑活动高度一致。一些早期研究发现，基于 RNN 或 Transformer 的语言模型可以解释大量语言相关脑区的活动，且模型的预测准确率与脑解码性能呈正相关。这些结果被视为支持大脑预测编码的证据。

需要强调的是，语言中的预测可能涉及不同层次：在语音层面，人类利用共现概率

预测下一个音素或重音；在词法层面，预测词的词形和词性；在句法层面，预测短语结构和句法角色；在语篇层面，预测话题发展或叙事结构。人脑可能在这些层次上并行生成预测，而当前神经语言模型主要专注于词序列层面的预测，忽视了语音和句法层面的单独预测。此外，语言模型中的误差信号仅用于训练过程，推理时不会在层间反馈，而预测编码理论强调误差信号自下而上传递、更新内部模型。早期 ERP 指标如 N400 与 P600 虽常被解释为语义或句法预测误差，但也有研究认为它们反映语义整合困难、冲突监控或重新分析等多种认知过程。因此，尽管神经语言模型的成功为预测编码提供了重要线索，我们仍需谨慎区分模型的预测目标与大脑真实的预测机制。

预测编码理论起源于贝叶斯大脑假说，将大脑视为一个生成模型，在每个时间步预测感官输入的概率分布，并通过最小化预测误差维持内外信息的一致性。在语言领域，这意味着不仅要预测即将出现的词，还要预测其词法形态、语法角色以及在语篇中的功能。例如，英语的限定词后可能接名词，而日语中的助词指示了不同的语法关系，大脑需要根据语言的语法规则调整其预测。不同语言之间的词序和结构差异也挑战了简单的线性预测框架，提示预测可能是多层次、动态调整的过程。

另一个相关的理论是生成式模型，它认为大脑在理解语言时会构建一个潜在的生成过程，形成关于事件、场景或对话的内在模型。这种生成式推理不仅包含预测，还涉及假设检验和信念更新。例如，在叙事理解中，听者会建立角色间的因果关系和动机结构，对即将发生的情节做出推断。当新信息与内在模型不符时，大脑将产生突出的预测误差信号，从而驱动理解的修正。神经语言模型多采用自回归训练，但真实生成过程可能依赖递归和层次化的预测，这也是模型尚未完全捕捉到的。

预测编码的思想早在上世纪九十年代在视觉系统提出，Rao 和 Ballard 提出了一个分层模型，上层神经元产生低层输入的预测，下层神经元反馈误差信号驱动上层更新。在听觉和语言研究中，类似的框架用于解释为什么语境越丰富，大脑的听觉皮层响应越弱：这种现象被解释为预测误差减少。该理论强调自上而下连接的作用，认为皮层层次之间通过反馈信号调节编码。但大脑皮层的解剖显示出复杂的双向连接，预测编码模型如何映射到这些解剖结构仍具争议。例如，颞上沟的层间连接既有前馈也有反馈，高层视皮层与耳蜗的反馈回路则难以用简单的预测误差解释。此外，语言中的冗余和多义性使得预测必须结合词义和语境，不同语言的后缀、语序自由度也影响预测策略。为了准确评估预测编码理论，研究需要在不同语言和任务上系统测量预测误差信号，并区分语法预测、语义预测和语用推理。

在预测编码的讨论中，还需区分不同层次的预测目标。生成式贝叶斯模型强调，大脑不仅预测即将到来的感官输入，还预测输入的原因。例如，听到特定音节后，大脑可能预测未来出现的语义类别或句法角色，而不仅是具体词形。语言理解涉及对外界事件和他

人意图的建模，这种隐式推理远超下一词预测。研究发现，当叙事违反人物动机或事件因果时，大脑的默认模式网络产生强烈反应，但这与词层面的 *surprisal* 无关。另一方面，预测并非总有利于理解；在幽默或修辞反转中，出乎意料的表达反而提高了记忆和理解。神经语言模型通常通过最大化下一个词的概率学习，但大脑在面对新颖和意料外的句子时可能主动抑制预测，以保持开放性。因此，未来实验应将预测难度、句式复杂度和语用意图作为独立维度操控，通过组合这些因素进一步测试预测编码假说的边界。此外，生成式模型与分布式语义模型并非对立，可通过变分自编码器等框架联合实现预测与解释；这些模型能同时学习生成过程和意义结构，为理解大脑如何整合预测与概念知识提供新的工具。

3.2 对预测编码证据的质疑

尽管语言模型在脑编码任务上的成功常被用作预测编码的证据，但这一推论遭到质疑。Antonello 与 Huth (2023) [1] 对神经语言模型与大脑匹配的机制进行了细致分析。他们发现，同一模型的不同层在解释大脑活动时表现迥异：用于词预测任务的高层隐藏状态并非最佳的神经预测器，相反，中间层的表示更能解释大脑数据。此外，他们提出了一个衡量模型在多种下游任务上迁移能力的“通用迁移性能”指标。研究发现，该指标与模型的脑编码性能同样相关，而与下一词预测性能的关系并不更强。因此，模型在大脑编码任务中的成功可能源自其学习到的丰富语言结构，而不仅仅是预测任务本身。

另一个支持这一观点的证据来自层级分析。对比模型不同层的表示，研究者发现中层能够捕获局部依存和语义结构，而高层则更专注于全局预测。在大脑编码任务中，中层表示的解释方差显著高于高层表示。如果预测编码是唯一关键因素，理论上最擅长预测任务的高层应该最能解释大脑活动，但事实恰恰相反。该研究因此提醒，不能简单把模型的预测目标与大脑的认知目标等同，也不能基于模型的预测准确率直接推出大脑采用预测机制。

在质疑预测编码假说的工作中，Antonello 与 Huth 对神经模型与大脑匹配度展开系统性检验。其数据集由五名健康成人组成，每名受试者在多次扫描中聆听约五小时的英语播客故事。研究者构建了 97 个特征空间，包括词向量、模型不同层的隐藏状态以及手工设计的语言学特征，并通过 Lanczos 插值将这些特征时间序列与 fMRI 采样率对齐。在回归模型中，他们分别比较不同特征对体素级响应的解释能力，结果显示，Transformer 模型的中层表示在预测脑活动时普遍优于后层和早层，且这种优势在不同被试和不同脑区中一致。

除了层级分析，研究者还引入通用迁移性能指标，用以衡量模型在情感分析、命名实

体识别、翻译等多任务上的表现。令人惊讶的是，该指标与脑编码性能的相关性与下一词预测相当，甚至在某些情况下更强，这表明模型的迁移能力可能比其预测能力更能反映大脑语言处理的本质。这一发现促使我们重新思考预测编码假说的适用范围：语言理解可能依赖多种学习机制，如统计学习、结构学习和语义推理，而不仅是简单的下一词预测。此外，作者强调，未来研究应通过实验操控模型的训练任务，如比较纯预测任务与多任务学习模型在脑编码中的差异，从而更严格地检验预测编码的贡献。

Antonello 与 Huth 的质疑工作不仅比较了模型层级的性能，还分析了由语言学家手工标注的特征，如词性、句法依存树、形态标签等。他们发现，这些传统特征单独时预测脑活动的能力远低于神经模型表示，但将其与模型嵌入组合可以略微提升性能，表明模型捕获的高级特征包含这些语言学信息。值得注意的是，他们使用约 5 小时的自然播客故事，总共约 45000 个单词，几乎涵盖多种主题和叙事风格。研究者利用卷积插值将单词特征对齐到 fMRI 时间点，并采用嵌套交叉验证确保统计可靠性。这样的严谨设计减少了过拟合风险，也为其他研究提供了基准。作者还指出，模型在任务外迁移性能上的表现和语法性判断、语义相似度等任务的相关性不高，这表明各任务衡量的语言能力不同，未来应构建多维评估指标综合评估模型与大脑的对应。

3.3 集成建模对预测性的支持

与上述质疑相对，另一系列研究采用大规模集成建模方法，通过比较不同模型、不同任务和不同数据集，系统检验模型性能与大脑匹配度之间的关系。Schrimpf 等（2021）[9] 汇集了 43 种语言模型，包括不同深度的 Transformer、循环神经网络（RNN）和静态词向量，并对比它们在多个神经和行为数据集上的表现。数据集包括人们阅读或听句子时的 fMRI 和 ECoG 信号以及反应时间等行为指标。结果表明，Transformer 模型显著优于 RNN 或静态嵌入，容量越大性能越好。更重要的是，模型的脑拟合度、行为拟合度与其下一词预测准确率之间存在显著正相关，而与其他语言任务（如句法分析、文本分类）无关。此外，未训练的 Transformer 也能解释部分大脑数据，表明模型架构对匹配度有基础性贡献。

综上，集成建模似乎支持预测编码：能够更好预测下一词的模型往往更能解释大脑数据。然而，这种关联并不能排除其他解释。例如，大型模型在训练过程中捕获了更丰富的语言统计规律，其表现优异可能来自于表示的复杂性而非预测任务本身。因此，预测与特征学习之间的贡献仍需通过控制实验加以区分。

集成建模工作汇聚了多个数据集，旨在探索模型性能与脑匹配度之间的普遍规律。Pereira2018 数据集包括 78 名参与者在阅读约 400 个句子时的 fMRI 信号，每个句子呈现

多次以提高信噪比。Fedorenko2016 数据集采用 ECoG，记录 12 名癫痫患者在阅读单词或短语时的皮层电活动，具有高时间分辨率，适合分析迅速的语音和语义过程。Blank2014 数据集让参与者聆听约五分钟的自然故事，捕捉更生态的语篇处理。通过在这些数据集上对 43 个模型的各层表示进行线性回归，研究者发现模型的脑拟合度与下一词预测准确率之间呈现强相关，且 Transformer 模型几乎达到了噪声上限。

深入分析表明，模型架构对大脑匹配度的贡献不可忽视。即便未经训练，Transformer 架构也能在一定程度上预测脑活动，这表明多层自注意结构本身具有与语言网络相似的组织方式。此外，模型的容量越大，匹配度越高，提示大脑语言系统可能利用高维表征整合丰富的语境信息。集成建模还强调，模型在其他语言任务（如问答、句法分析）上的性能与脑拟合度几乎不相关，说明预测任务在当前模型框架中仍然是最能反映大脑语言处理的代理任务。然而，这并不意味着大脑只做预测，而可能是目前的预测任务同时涵盖了语义、语法等多维信息，所以表现出较强的相关性。

Schrimpf 等人的集成建模研究除了下一词预测任务，还评估了模型在机器翻译、问答、句法分析等任务上的性能。他们发现这些任务的准确率与脑拟合度之间相关性较低，这意味着大脑理解语言可能不依赖模型在某些人工任务上的表现。研究还发现，不同受试者、不同刺激材料和不同脑区之间的匹配度差异较小，说明某些规律具有普遍性。他们还使用双任务对比，比较随机初始化模型、预训练模型和经过微调的模型，结果显示预训练是获得高脑拟合度的关键，而微调对特定任务并不能显著提升脑对齐。这一观察支持了模型通过大规模无监督学习捕获人类语言统计结构的重要性，同时也提示未来模型设计应优先考虑预训练阶段的任务和数据多样性。

集成建模之所以能够揭示模型性能与脑拟合度的关系，得益于对多样数据集和多重评测指标的系统整合。Schrimpf 等人在比较 43 个语言模型时，不仅考虑模型的下一词预测准确率，还综合了模型规模、层数、训练语料大小以及是否使用双向或自回归架构。他们发现模型容量与脑拟合度呈非线性关系，层数增加初期可显著提升预测能力，但超过一定深度后收益递减。数据量方面，模型在数十亿词语料上训练可以捕获更丰富的语义和句法统计信息，但超过某一阈值后提升有限。研究还比较了微调模型在特定任务（如翻译、问答）上的表现，结果显示针对下游任务的微调有时会降低脑拟合度，可能因模型过拟合任务数据而破坏其通用表征。这提示我们，在构建大脑对齐的语言模型时，应平衡任务适应与通用语言知识的保持。此外，集成建模涵盖的神经数据主要来自英语，未来应引入其他语言和文化的脑数据，将模型的普适性检验扩展到更广泛的语言生态。

4

语义系统的脑映射与分布式表征

4.1 数据驱动绘制语义地图

在自然语言理解中，大脑对语义信息的表示是分布式的。Huth 等（2016）[4] 通过让受试者聆听长达两个小时的自然故事，利用词共现构建的 985 维词嵌入和体素级正则化回归，绘制了语义系统的高分辨率地图。模型经交叉验证可预测新故事的 fMRI 响应，说明嵌入捕获了稳定的语义特征。通过对模型权重进行主成分分析，他们提取出四个跨受试者共享的主要语义维度，并使用 PrAGMATIC 算法将这些维度投影到皮层，发现左半球约含 77 个语义区域，右半球约 63 个。这些区域不仅涉及传统语言区，还扩展到顶叶皮层（LPC）、内侧顶叶（MPC）和前额叶的默认模式网络（DMN）区域，其中中心区域偏向社会和人物概念，外周区域偏向数字、视觉或触觉概念。这一发现打破了仅有左侧优势的传统观点，揭示语义系统在左右半球间较为对称。

除绘制地图本身外，Huth 等还详细描述了模型的训练和评估流程。他们构建了 985 维语义特征矩阵，其中每一维度表示英语词与语料中其他词的共现概率。为了消除低级听觉因素，模型在回归时加入词率、音素率和声学特征作为协变量。在训练阶段，研究者利用 10 折交叉验证评估模型对新故事的预测能力，保证了结果不依赖特定刺激。这种严格的验证使绘制出的语义地图具有可重复性和泛化性。通过观察各语义维度在皮层表面的分布，他们发现概念的抽象程度呈后—前梯度：背侧和后侧区域偏向具体感官相关概念，如动作、视觉和听觉；腹侧和前侧区域则偏向抽象、社交和情感概念。该渐变跨越颞叶、顶叶和前额叶，反映大脑可能按抽象度或表象类型组织语义信息。作者建议，这一渐变与从知觉到抽象思维的连续加工路线相一致，为理解概念结构提供了新的神经学证据。

这种数据驱动的语义映射方法具有多重意义。首先，它利用自然故事这一生态刺激，克服了传统实验采用单词或短语的限制，展示出在真实语境下绘制语义地图的可行性。其次，语义地图可作为参照框架，便于不同研究之间比较语义表征位置。作者指出，未来研究需改进区域划分算法，兼顾离散区域与功能渐变。此外，通过跨语言和跨文化采样，可以检验语义系统的普遍性和可变性。

基于自然故事的语义映射提供了更全面的视角。一些后续研究将 Huth 等的方法扩展到其他语言，如西班牙语、法语和中文，发现语义地图的宏观结构具有高度一致性，这表明大脑语义系统的组织具有跨语言普遍性。然而，在某些文化特定概念上，如颜色词、

亲属称谓或食物名称，不同语言的激活模式存在细微差异，这可能与语言中相关类别的词汇丰富度或社会文化重要性有关。通过比较不同语言的语义地图，研究者可以揭示概念表示的文化可塑性及其神经基础。

语义地图还揭示了默认模式网络在语义处理中的核心地位。DMN 包含的角回、后扣带皮层和内侧前额叶长期以来被认为参与自发思维、内省和记忆检索。Huth 等的结果显示，这些区域也积极参与语义加工，尤其是涉及社会、情感和思维推理的概念。这一发现使我们重新审视 DMN 的功能定位，提示它可能不是专门处理“与任务无关”的思维，而是在语义推理过程中发挥中枢作用。语义地图的结果还打破了传统认为左半球主导语言的观点，揭示两半球在语义任务上的对称性。不过，由于 fMRI 在前颞叶的信号噪声较大，某些语义区域可能仍未被发现，因此需结合 ECoG 或高场强 fMRI 提高空间分辨率。

在对语义地图的后续分析中，研究者进一步解析四个主要语义维度分别对应哪些概念群。例如，第一个维度从有生命的生物到无生命物体，反映动—静连续体；第二个维度从社会交往到工具使用；第三个维度从视觉场景到感知属性；第四个维度从数量与空间到情感与心理状态。通过比较这些维度在皮层上的渐变，发现背侧区域偏向具体感官经验，腹侧区域偏向抽象社会知识。实验还比较了不同故事段落中语义向量的变化，结果显示语义维度的激活模式随故事推进而动态演化，表明语义处理具有时间依赖性而非静态。作者提出，可以将语义地图作为生成刺激的指南，选择刺激中激活特定区域的词语或句子，研究这些区域对语义推理的因果作用。

语义地图不仅揭示出概念在皮层表面的分布，还呈现出随着故事语境变化而动态更新的模式。后续研究利用滑动窗口技术，分析语义维度的激活随时间的变化，发现故事高潮时期社会和情感维度的激活显著增加，而叙述背景期则更多激活物体和场景维度。这表明大脑语义表示具有时间敏感性，会根据故事进程调整重点。另一些研究通过比较不同叙事体裁，如对话、新闻报道和诗歌，发现诗歌中的抽象情感维度激活更强，而对话和新闻则更依赖社会互动维度。除此之外，跨语境分析显示，同一概念在不同故事中的激活模式可能不同，反映语义表示与篇章背景的耦合。未来可以结合自然语言生成模型，控制故事内容和风格，系统探索语义表示的动态调节。进一步地，利用联结梯度分析，可以描绘语义网络在皮层内部的连续变化，揭示从感觉相关区域到抽象思维区域的渐变。这种多维度、多时间尺度的语义地图将为理解语言语义的动态生成提供新的视角。

4.2 语义类别与关系的分布式编码

在对语义地图的进一步探索中，Zhang 等（2020）[12] 让受试者听 11 小时故事，建立体素级编码模型，预测数千个单词的脑响应。他们发现，大脑并不以离散模块表示不同

语义类别，而是通过广泛重叠的区域同时编码多种类别。例如，工具类概念在左侧顶下小叶、后中颞回和颞上回均有表示；交流与情感等抽象概念则在右前颞区和顶叶更为明显。通过分析词汇的具体性，研究者观察到左半球偏向具体、感官相关概念，而右半球更偏向抽象、内省相关概念。这一左右半球差异揭示语义系统在处理不同类型概念时的功能特化。

不仅语义类别，概念之间的语义关系也可以映射到皮层。*Zhang* 等利用词向量差表示语义关系，例如整体—部分、类—属、对象—属性等，并构建关系编码模型。他们发现，“整体—部分”关系在默认模式网络区域（如角回、后扣带皮层）呈现明显激活，而前顶叶注意网络呈现抑制；其他关系则在不同网络中表现不同的激活抑制模式。这种共同激活与抑制的模式表明，大脑通过协同的功能网络而非独立区域编码语义推理。作者还发现语义关系的网络模式与语义类别脱钩，例如“手—手指”和“动物—动物园”属于不同类别但具有相似的关系模式。这表明大脑可能存在专门处理抽象关系的网络，与 DMN 中的思维漫游或内省功能相联系。

值得进一步说明的是，*Zhang* 等划分的九个语义类别包含工具、人类、植物、动物、地点、交流、情感、变化和数量等，每个类别又由数百个单词组成。对这些类别的分析显示，具体概念（如“锤子”“狗”“苹果”）在多感官和运动相关区域呈现较强激活，而抽象概念（如“自由”“希望”“交流”）则在前额叶和顶叶默认模式网络表现更强。语义关系方面，除了整体—部分、类—属和对象—属性，研究者还考虑了名词与动作之间的关系、时间关联、空间关联和事件因果关系。他们利用 SemEval-2012 评测集中的句子对构建差向量，并利用这些向量预测脑响应。结果表明，不同关系在皮层上呈现高度一致的空间模式，说明大脑可能通过共享的网络处理各类关系推理，而不是为每种关系单独配置区域。这一发现拓展了我们对语义系统功能的理解，表明抽象关系加工与默认模式网络的内在思维密切相关，并可能涉及对情境和情感的整合。

Zhang 等进一步分析了不同类别词汇在皮层中的精细分布。例如，人类和动物概念不仅在后颞与顶叶区域活跃，还在视觉皮层的被动激活区出现，可能与想象或回忆相关；情感和交流类词汇在角回、内侧前额叶等 DMN 区域更强，说明这些抽象概念与自我反省和社会认知密切相关。研究者还比较了不同关系类型，如“物品与使用者”“因果与结果”“部分与整体”，发现关系向量激活模式的相似性与关系的逻辑结构相关。例如，因果关系与时间关系的皮层模式相近，体现故事中事件连贯性在脑中的共通表示；而反义关系激活的网络更分散，可能需要更多注意和工作记忆资源。这些发现提示语义关系不仅是词义差向量，在大脑中也体现为跨网络的协同模式。

4.3 地图与网络的意义

语义地图和语义关系研究表明，语义系统既有局部分区又存在跨区域的连续功能梯度。Huth 等的 PrAGMATIC 算法划分出多个语义区域，但假设每个区域内部同质，这难以捕捉某些功能渐变。未来需要发展既能识别离散区域又能描述功能渐变的模型，如连通性梯度分析。Zhang 等的研究强调语义关系的网络化特点，通过默认模式网络与注意网络的协同活动编码抽象关系。这些发现提示，语义认知不仅依赖单个区域的选择性，也依赖跨网络的动态交互，理解语言中的推理和抽象思维需从网络视角入手。

通过结合皮层连接信息，研究者建议语义系统可划分为若干功能网络：中心的 DMN 支持抽象概念和情景推理，背侧注意网络支持概念的检索和选择，腹侧语义网络支持具体物体和动作信息。语义关系的编码往往跨越这些网络，例如“部分—整体”关系需要同时激活 DMN（处理整体概念）和抑制顶叶注意网络（抑制无关信息）。这表明语义处理可能涉及网络间的抑制与兴奋平衡。未来需要采用动态图模型或有效连接分析，理解在语义推理过程中网络交互的因果顺序。此外，语义网络与其他认知网络如工作记忆、情绪和奖励系统的交互也值得探索，因为实际语境中的语言理解往往伴随情感体验和行动决策。

5

学习机制与模型：自监督语音与 Transformer

5.1 自监督语音模型 wav2vec 2.0

人工语音学习通常不依赖大规模标注，因此自监督语音模型能更贴近人类语言获得的条件。Millet 等 (2023) [6] 考察了自监督语音模型 wav2vec 2.0[2] 与大脑活动的对应性。wav2vec 2.0 由三个主要部分组成：一个由七个卷积块构成的特征编码器，将原始 16 kHz 语音转换为低维潜在表示；一个量化模块，将连续表示映射为有限的离散符号词典；以及一个 12 层的 Transformer 上下文网络，通过自注意机制整合长距离信息。模型的自监督训练目标是预测被掩码帧的离散表示，训练引入对比损失和多样性损失，使模型既能依赖上下文，又能充分利用离散向量。

作者比较了随机初始化、自监督训练、监督训练以及跨语言训练的模型，并以这些模型的层级表示为特征，建立线性回归预测多个受试者聆听有声书时的 fMRI 响应。研究发现，自监督模型在大约 600 小时未标记语音上即可学习出与人脑相似的表示，模型的卷积层、量化层和 Transformer 层分别对应人类语音皮层的不同处理阶段。中层表示在听觉皮层和颞上沟表现最佳，后层 Transformer 表示则更符合语言皮层的反应。这些趋势在法语和普通话受试者中同样显现，表明自监督模型捕获了跨语言的声学与语音规律。行为实验结果也显示，模型的层级专化与人类语音辨别任务表现一致。

Millet 等的分析还指出，自监督模型的成功不仅依赖大规模训练数据，还受数据多样性和音频质量的影响。他们训练了不同尺寸的模型，发现 50 小时的数据即可学习基本的声学表示，但在更高语音层级（如音节、韵律）上需要数百小时的数据才能达到与大脑类似的层级专化。此外，作者比较了在单个语言和混合语言数据上训练的模型，发现跨语言训练的模型能更好地泛化到新语言并保持与大脑的高匹配度。这意味着，模型在学习语音规律时可能捕获了普遍的声学约束而非特定语言的词汇规则。他们还探索了监督学习模型（如声学—词法一体化模型），发现这类模型在高层表示上过度偏向目标任务（如字符识别），与大脑的匹配度反而下降。因此，自监督学习在模拟人类语言习得方面具有优势，未来可将其推广到更复杂的音系与语调结构，并与基于预测的语言模型整合。

自监督语音模型的优势不仅在于不依赖人工标签，还在于能够捕获语言通用的韵律和声学特征。wav2vec 2.0 的特征编码器由七个卷积模块组成，每个模块通过步长和卷积核大小控制时间分辨率和特征维度。在训练过程中，模型通过对比学习鼓励不同语音片段具有辨别性，同时利用量化模块构建离散的代码本，使连续语音表示能够映射到有限的符号集合。这些符号类似于音素或音节的抽象单位，为后续上下文网络提供清晰的离散输入。研究表明，自监督模型的不同层在处理语音的各个阶段表现出特定功能：高层卷积层对短时间帧的频谱特征敏感，中层表示音节和共振峰结构，后层 Transformer 捕获语调、词汇和发音风格。相比之下，监督训练的声学模型往往过度优化于特定标注任务，如语音识别或声学模型，从而导致其内部表示失去通用性，无法很好地解释大脑数据。通过在多种语言和方言上训练自监督模型，可以构建语言无关的声学基底，然后在此基础上微调特定语言的声学和语音特征。这一策略有望缩小模型与大脑在语音处理上的差距，特别是在处理口音、方言和情绪时。

5.2 Transformer 的计算与功能专化

Transformer 通过多层的自注意机制处理序列信息。此前研究多关注模型的隐向量嵌入，而 Kumar 等（2024）[5] 直接分析了注意头执行的变换，即每个注意头如何更新词

表示。他们将模型的变换分解为每个头的线性变换，并使用自然听故事的 fMRI 数据评估其预测能力。结果显示，注意头变换在语言皮层大部分区域能解释大量方差，且在后颞区显著优于传统语言学特征（如词性或句法依存关系）。不同层的头展示出渐进式的功能梯度：早层、短距离回溯的头权重在后侧颞区更高，而高层、长距离回溯的头则在前额与前额叶区域占优势。这种梯度与语言处理从局部到全局的层次结构一致。

此外，作者发现某些特定头对特定语法依存关系（如补语从句、直接宾语）有选择性，在后颞区表现尤为明显。值得注意的是，在角回等高级语义区域，非上下文嵌入的预测能力反而超过了变换，这表明这些区域可能整合全局语义内容而非依赖局部注意。作者强调，头部变换只是模型的线性近似，不代表大脑的真实计算，但这种方法提供了更加细粒度的功能对齐视角。他们建议未来探索瓶颈 Transformer 等新架构，引入声学特征，并结合行为和语言任务，通过梯度约束和多任务学习进一步靠近大脑机制。

Transformer 的自注意框架不仅提供了线性近似，还可能在不同层捕获句法层次、共指链和话语主题等丰富信息。近期分析发现，注意头不仅沿层次形成回溯距离的梯度，还沿功能类别发生分化。例如部分头专注于主谓一致、名词短语结构，另一些头捕获语气助词和焦点标记，反映语用提示。更重要的是，头与头之间存在协调机制：在处理长句或嵌套从句时，前层的局部头预先筛选相关修饰语，后层的全局头再根据语篇结构整合信息。实验表明，当人为阻塞关键注意头时，模型在脑编码任务上的预测性能会显著下降，这支持了头之间协同工作的观点。进一步的诊断显示，注意分布与人类阅读的眼动轨迹相关，高层注意与长距离回溯的阅读阶段吻合，说明模型学习到了与人类相似的注意模式。

未来研究可以在模型中纳入生物约束，如稀疏连接和时间常数，使注意机制更贴近突触动力学。例如可以采用动态稀疏化策略限制每层仅激活少数关键头，并让头的权重随时间衰减，模拟大脑对旧信息的遗忘。此外，可以结合符号化模块，将注意头分配给特定的句法功能或语义角色，实现可解释性更强的混合模型。鉴于不同头在不同语言和任务上的作用存在差异，跨语言分析注意模式将有助于识别普遍与特定的注意策略。通过这些改进，我们期望 Transformer 不仅在行为表现上接近人类，还能在内部计算上更符合大脑的层次组织。

此外，注意头的输出可视为一系列线性变换，将当前词向量映射到多个上下文子空间。研究显示，这些变换在处理隐喻、修辞重复和语调变化时表现出特定的模式，表明模型内部对语义和语用的编码比单纯词级向量更丰富。一些工作提出将注意机制与可微栈或记忆模块结合，捕捉跨句段乃至跨文档的长期依存。将这些结构嵌入语言模型，或可更好模拟大脑在保持情节线索和角色身份时的持久性。由于自注意计算的复杂度随序列长度平方增长，与大脑的时间和能量约束不符，近年出现的稀疏或线性注意机制能够在提高效率的同时保留建模能力。这些机制也许更贴近大脑的局部连接和长距离白质束传

导特性。对 Transformer 计算的深入理解和重构，将推动下一代高效且生理友好的模型，为解释大脑语言加工提供更多线索。

除了自监督和预测式学习机制外，最近的研究还探索了对比学习、变换器自回归与去噪自编码器结合的框架，以更全面地模拟语言习得。例如，带噪预训练的文本编码器通过随机遮挡句子片段并要求模型复原原句，在学习恢复局部信息的同时，还培养了对全局语义和句法的理解。这种任务结合了预测和重建两个目标，既包含自底向上的模式提取，也包含自顶向下的生成假设。实验证明，这类模型在脑编码任务中表现优异，尤其在预测内侧前额叶和角回等高层语义区域的活动时超越传统自回归模型。另一些研究引入了元学习和持续学习框架，使模型能够在不断变化的语料环境中适应新的语言模式，同时保留旧知识。这与大脑在终生学习中不断更新语义地图的过程类似。将这些先进的学习策略与神经数据对比，不仅可以评估它们的生物合理性，还可以揭示大脑可能采用的学习策略组合。

6

组合语义与 supra-word 表征

语言理解不仅涉及单词的意义，还需要基于上下文组合词汇产生超越字面含义的“超词”(supra-word) 意义。Toneva 等 (2022) [11] 提出了一种数据驱动方法，利用 ELMo 模型 [7] 的前向 LSTM 隐状态构建上下文嵌入，并通过线性回归消除单独词义的贡献，得到仅包含组合信息的残差嵌入。这种嵌入能捕捉隐含意义，例如 “Mary finished the apple” 中隐含的“吃完苹果”或“绿色香蕉”表示“未熟香蕉”，也可表示新语义组合。

将 supra-word 嵌入作为预测变量，研究者用其解释 fMRI 和 MEG 数据，发现经典词汇枢纽，如颞上回后部和颞下回，同样维护组合语义。此外，前颞叶也对 supra-word 嵌入敏感，说明词汇与组合语义在大脑中共享神经基础，需要前后颞区域协同维持。然而，他们在 MEG 数据中未检测到 supra-word 表征。这提示 supra-word 意义可能通过持续的、非同步的神经活动体现，而 MEG 对同步活动敏感，难以捕捉这种慢速信号。该结果强调不同脑成像技术对语义组合的敏感性不同，未来需要结合低频功率或相位同步等指标并使用多模态数据共同分析。

值得一提的是，supra-word 嵌入不仅揭示了组合语义的存在，还证明了分布式向量可以通过残差运算表示复杂的语义组合。这种方法与传统基于语法树的组合不同，它不依赖手工定义的规则，而是由模型在大量语料中学习到的统计规律。研究人员发现，将

supra-word 嵌入与词级嵌入结合，可以更准确地预测读者对隐含意义的理解程度。这一发现为改进语言模型提供了启示：未来可尝试不同的向量运算（如加权平均、向量差、张量积）以及递归或注意机制，来模拟人类如何积累和整合组合意义。此外，在神经数据中考察 supra-word 作用的时间动态是一个开放问题，未来可通过在 MEG 或 ECoG 数据中分析低频功率或相位同步来捕捉持续性组合信号，并结合工作记忆和注意任务，探讨语义组合与认知资源的关系。

除了 ELMo 生成的残差嵌入外，后续研究还探索了多种模型和运算来捕捉组合语义。例如，通过比较 GPT-2[8] 或 BERT[3] 的上下文嵌入与各单词嵌入的向量差，可以获取另一种 supra-word 表征，这些表征在脑编码任务中的表现与 ELMo 残差相当。另一个方向是使用张量积或基于张量神经网络的方法，将两个词向量结合为高阶张量，从而显式表示交互项。虽然这些方法在计算上更昂贵，但它们提供了更丰富的组合信息。研究还发现，组合语义的表征不局限于双词短语，复杂句子中的嵌套关系和修饰语也可以通过递归地应用残差运算来分解。实验表明，这种递归嵌套的 supra-word 嵌入能够更好地预测人类对歧义句子的理解方式，说明模型对多义性消解有所把握。

组合语义不仅依赖词汇的线性组合，还与语法结构和语用知识紧密相关。在一些语言中，形态变化（如词序倒装、语气助词）会改变句子意味和语域，模型需要学习这些语法标记如何影响组合意义。此外，跨语言比较发现汉语、日语等语言的复合词和固定搭配大量依赖语法化的组合规则，这使得 supra-word 表征在这些语言中可能具有不同的统计特征。未来研究应扩展到不同语言的组合语义，检验模型能否捕获这些语言特有的组合规律。为了更精确地测量 supra-word 处理的神经时间动态，可以利用 ECoG 或高时间分辨率 fMRI，结合语音停顿和语调变化，分析大脑如何在听话者缓慢而连续的输入中形成语义组合。总之，supra-word 研究为我们揭示了词汇组合的复杂性，提醒语言理解是一个多层次、多维度的动态过程，需要模型和神经数据同时考虑语法、语义和语用因素。

7 逆向工程：从神经信号重构连续语义

除了“编码”（从刺激预测脑反应），近年来研究也开始系统推进“解码”（从脑反应重构刺激）与“逆向工程”。Tang 等（2023）[10] 展示了利用非侵入式 fMRI 在连续叙事场景下重构语义内容的可行性：模型并非逐字逐句复原原文，而是重构与原叙事在语义上高度相近的表达。这一结果提示，大脑在自然语境下的表征更接近“意义层级”的压

缩，而不是对词形的逐点记录。

在方法上，解码任务常被形式化为寻找最可能的文本序列 S :

$$\hat{S} = \arg \max_S P(S | R) \propto \arg \max_S P(R | S) P(S),$$

其中 R 表示脑反应（如 fMRI 体素时间序列）， $P(R | S)$ 对应“编码模型”，而 $P(S)$ 则提供语言先验（可由语言模型给出）。在实际求解中，研究者通常在候选空间内采用波束搜索（beam search）等策略，平衡“脑一致性”与“语言可行性”。这类工作也带来重要的科学与伦理问题：它既为检验语义系统的分布式表征提供了新的量化指标，也推动了对脑数据隐私与可解释性的讨论。

8

方法论：编码模型、评价指标与统计检验

8.1 编码模型：岭回归与分带岭回归

当以高维模型表征作为特征时，脑编码模型往往面临 $P \gg N$ 的病态问题，因此线性回归通常配合正则化使用。最常见的形式是岭回归：

$$\hat{\beta} = (X^\top X + \lambda I)^{-1} X^\top y,$$

其中 X 为特征矩阵、 y 为神经响应、 λ 为正则化系数。在特征由多子空间构成（例如声学特征与语义特征）时，还可使用分带岭回归（banded ridge regression），为不同子空间分配不同的正则化强度，从而进行方差分解并更清晰地区分不同信息源的独立贡献。

8.2 噪声天花板与显著性检验

由于神经测量本身存在噪声，模型预测的相关性上限受到“噪声天花板”（noise ceiling）约束。实践中常以重复试次的一致性或跨被试一致性估计可解释方差，从而避免对模型性能的过度解读。对单体素或单通道的显著性判断，则常结合置换检验（permutation test）构建零分布，以控制多重比较并提升结论的稳健性。

跨语言与跨模态比较

跨语言比较有助于检验模型与大脑匹配的普适性。Millet 等纳入英语、法语和普通话受试者，发现 wav2vec 2.0 自监督模型在不同语言中的层次映射非常一致：卷积层对应基本声学特征，中层对应语音特定特征，后期 Transformer 层对应语言特定信息。这一跨语言一致性表明，自监督模型捕捉了普遍的声学和语音规律，并且大脑对不同语言的加工共享相似的功能层级。

跨语言研究还应考虑语言类型学和文化差异。一些初步工作将语义地图方法应用于西班牙语、日语等非印欧语系，发现基本的语义区域位置类似，但某些文化特定概念（如礼貌等级、敬语、宗教词汇）在皮层中的激活强度存在差异。这可能反映不同语言在语义编码上的策略，以及语言经验对神经表征的塑造。此外，跨语言比较可以揭示不同书写系统对语义加工的影响。例如，表意文字（如汉字）阅读者更依赖视觉形状与语义联想，而表音文字阅读者更依赖声音规则。大脑在这两类文字的处理过程中激活的视觉和语音区域有所不同，这些差异应在模型评估中加以考虑。未来扩展到双语者和方言使用者，将有助于理解语言经验对语义系统的可塑性和适应性。

跨模态比较则揭示不同成像技术对语言处理的敏感性差异。Schrimpf 等的集成建模同时分析了 fMRI 和 ECoG 数据，发现 Transformer 模型对两种模态的预测能力高度一致。然而，Toneva 等发现 MEG 数据无法检测 supra-word 表征。这些对比强调，需要结合多模态数据以捕捉大脑语言处理的不同时间和空间尺度。例如，fMRI 能捕捉慢速血氧反应，适合发现持续性语义信息，而 MEG 则对快速电同步更敏感，适合捕捉即时处理。这些差异需在模型与大脑对齐时仔细考虑。

跨语言研究还强调了语言类型学的多样性对语义系统的塑造作用。黏着语（如土耳其语、芬兰语）通过在一个词上附加多个语素表达语法关系，大脑可能利用这些形态学线索在早期阶段预测词干和附加成分的组合。屈折语（如拉丁语、俄语）则通过词形变化表示时态和格标记，需要跨更长距离的依存关系解析。孤立语（如越南语、泰语）依靠语序和功能词表达语法，使得句子结构的预测依赖于对短语层级的掌握。这些差异影响了语言模型和大脑在预测下一词时使用的策略，也要求我们在模型评估中纳入多样的语言。

文化因素也通过语言使用频率、隐喻习惯和礼貌策略等影响语义系统。例如，某些文化中对亲属关系有精细的词汇区分，而在另一些文化中对颜色或味觉的词汇更为丰富。研究者在跨文化采样时发现，与家庭和情感相关的词在东亚文化中引发的皮层激活更为

广泛，而与个人独立性相关的词在西方文化的前额叶激活更强。这些差异提醒我们，语义地图具有可塑性，会随着语言环境和社会规范调整。未来应在不同文化群体中重复语义映射和脑编码实验，构建跨文化的语义参考图谱。

跨模态比较的意义不仅在于技术差异，还在于不同信号反映的神经营过程各有侧重。ECoG 捕捉皮层表面电位，能实时反映快速同步活动，适合研究词汇和音素边界的瞬时处理；fMRI 捕捉血氧变化，反映几秒内的总体活动，适合研究持续的语义和情节加工；MEG 介于二者之间，反映大规模神经群体的同步，但对深层结构的敏感性较低。通过在同一受试者身上同时采集或跨实验结合这些数据，可以构建时间分辨率和空间分辨率兼顾的动态语义模型。例如，可先利用 MEG 确定 supra-word 组合发生的时间窗，再利用 fMRI 确定具体位置。跨模态融合将带来更全面的语言加工图景，促进理论与模型的发展。

跨语言语义映射的进一步研究需要涵盖非印欧语言群体以及低资源语言，如非洲的班图语系、美洲的纳瓦荷语或澳大利亚的皮京语。这些语言在语音、形态和句法结构上具有独特特征，如点击音、序列化动词或双数标记，可能导致不同的语义组织模式。通过将这些语言纳入语义地图项目，可以检验语义系统的普遍性，并为语言多样性保护提供科学依据。此外，随着语言接触和全球化的深入，多语言者的大脑可能展现出更加灵活的语义网络和预测策略。研究显示，精通多种语言的人在切换语言时能够快速调整语义激活模式，这反映出一种在语义空间中的动态抑制与增强机制。将这种多语灵活性纳入神经模型，不仅有助于理解双语脑如何管理多个词汇和语法系统，也能启发开发能够动态切换语境的人工模型。未来的跨语言研究应将样本扩展到不同的社会群体，包含方言和混合语言，如新加坡英语和斯普兰语，从而捕获语言演化和创新对语义系统的影响。

10

讨论：综合视角与未来方向

10.1 预测还是特征学习？

关于神经语言模型在大脑编码任务中成功的解释存在两种观点。一方面，集成建模显示模型的脑拟合度与下一词预测准确率高度相关。这一结果常被解读为大脑优化于预测未来输入。另一方面，层级分析与通用迁移指标则指出，模型的预测能力不一定决定脑编码性能，中间层或迁移能力更能解释大脑数据。因此，我们需要超越简单的二分法，认识到预测和特征学习可能共同作用：大脑在处理语言时既利用预测生成候选，还依赖丰

富的统计结构进行解析。未来研究可以通过操纵模型的训练目标（例如仅训练下一词预测 vs. 联合其他任务）来评估这些因素在脑编码中的相对贡献。

10.2 分布式语义系统与网络化视角

语义地图和语义关系研究揭示，大脑语义系统并非由少数专门区域组成，而是由广泛的分布式网络构成。PrAGMATIC 地图揭示 DMN 内有多个语义亚区，且左右半球分布对称。Zhang 等进一步发现语义关系通过 DMN 的激活与前顶叶注意网络的抑制共同编码，说明抽象推理依赖网络的动态交互。理解语义系统需要关注这些网络的连接模式和时空动态，而不仅是个别皮层区域的选择性。未来可以利用功能连接分析或图论方法，探索语义网络在不同任务和状态下的重构与调节。

10.3 模型特性与大脑对应

神经模型的学习机制和架构对与大脑的对应性具有重要影响。自监督语音模型在少量未标记数据上就能学得与大脑相似的层级表示，强调自监督和对比学习的价值。Transformer 的注意头分析提供了更细粒度的功能对齐视角，揭示不同层和头在皮层中的对应关系。这些方法说明，大脑和模型可能共享某些计算原则，如局部与全局整合。但目前模型仍缺乏对生理约束的考虑。加入时间连续性、能量限制、发展过程等约束，可能使模型更符合大脑。

10.4 组合语义的新挑战

supra-word 研究指出，组合语义在大脑中与词汇语义共享神经基础但依赖不同的维持机制。当前的语言模型通常通过固定窗口产生静态嵌入，难以捕捉组合语义的动态特性。未来应探索结构化或递归模型，在模型中显式实现组合运算，并比较不同组合算法对脑数据的预测能力。此外，需要在 MEG 和 ECoG 中寻找捕捉组合语义的合适信号，例如低频功率或相位同步。

10.5 跨文化与多语言的扩展

现有研究主要使用英语和西方语言，受试者背景也相对单一。Huth 等指出不同个体间语义地图的相似性可能源自共同的生活经验；Zhang 等则发现抽象概念与右半球关联的强度可能受文化差异影响。因此，未来需要在不同文化和语言环境中采集数据，比较语

义系统的共性与差异。研究应扩展到儿童、双语者或方言使用者，以揭示语义系统的发育和可塑性。这将帮助我们理解语言经验与神经组织如何交互，促进构建具备跨语言普遍性的模型。

讨论未来方向时，还需考虑神经语言模型与其他认知系统的交互。语言理解往往伴随记忆检索、情感评价和动作规划等过程，语义系统必须与海马回、杏仁核以及额顶网络协同工作。现有模型多数仅处理语言输入，缺乏与视觉和情感系统的互动。将视觉和情感通道融入模型，可以模拟故事理解中的场景想象和情绪反应。例如，电影描述不仅包含语言，还伴随画面和音乐，这些多模态刺激触发更复杂的语义及情感网络。开发能够同时处理文本、图像和声音的多模态模型，并用其预测多模态脑数据，将为全面理解自然语境下的语言加工奠定基础。

另一个重要方向是发展因果推断的方法。目前多数研究基于相关分析，难以确定模型表示对大脑活动的因果作用。结合经颅磁刺激、脑损伤研究或神经反馈，可以测试某些模型表征是否必要或充分。例如，当模型预测出高语义突发点时，相关脑区是否必然会出现因果响应？机器学习中的可解释性工具（如特征重要性和层次可视化）也可与神经干预结合，验证特定特征在语义加工中的作用。通过这些方法，我们可以揭示模型与大脑之间的因果联系，而不仅是相关性。

此外，伦理和公平问题应贯穿研究全过程。语义系统与价值观和社会经验密切相关，训练数据中的偏见可能导致模型学习到有害的语义关联，例如性别刻板印象或种族偏见。在比较不同文化和语言时，也必须尊重多样性并避免文化中心主义。开发公平、透明且可解释的语言模型，并理解其与大脑语义系统的异同，将有助于构建包容、安全的人工智能系统。

另一个值得关注的领域是儿童语言学习与成人语言理解的差异。婴幼儿在缺乏明确监督的情况下，通过与环境互动自然习得语言，其学习过程可能更依赖于语音和语义的统计共生，而非明确的预测。研究表明，儿童的大脑在处理语言时更依赖音节层级和韵律模式，而成人则能利用句法和语义进行更高级的预测。当前神经语言模型主要基于成人语言语料训练，忽视了语言发展阶段的差异。未来可以通过训练自监督模型在儿童对话语料上，模拟语言的发展过程，并与不同年龄段的脑成像数据比较，揭示语言系统的成熟轨迹。此外，还应关注老年人的语言加工变化，研究老化和神经退行性疾病如何影响语义网络和预测机制。通过跨年龄的比较，我们可以建立更加全面的语言认知模型，为教育和康复提供科学依据。

从更广阔的角度看，语言研究与社会科学、哲学和人工智能伦理密切相关。语言不仅是沟通工具，也是思想和文化的载体。神经语言模型的广泛应用有可能影响公共舆论和社会认知结构，因此研究者需要审慎评估模型的社会影响。例如，语言模型生成的内容可

能强化现有偏见或创造新的误解；模型的预测机制可能在某些文化中被误解为超越或取代人类思考。通过与哲学家、社会学家和伦理学家的合作，可以更深入理解语义系统与社会结构的相互作用，并制定适合不同文化环境的技术规范。科学家还应通过科普教育，让公众理解神经语言模型的能力与局限，建立合理期待，避免神话化技术。综观全局，语言的神经与计算研究既是科学探索，也是社会实践的一部分，必须兼顾科学价值和社会责任。

11

第二部分：实验设计、方法与结果

12

背景、目标与本报告范围

本项目采用自然故事听觉范式：被试连续听取长时语音刺激，研究者同时记录全脑 fMRI。与经典的离散刺激范式相比，这类数据的时间结构更接近真实语言理解过程，刺激在声学层面随时间快速变化，而语义与叙事层面的信息则跨句、跨段累积。对于编码建模而言，刺激的这种层级结构意味着两件事。第一，刺激特征必须被严格对齐到 fMRI 的采样时刻 (TR)，否则编码模型的输入输出不在同一时间轴上，任何“模型表征与脑表征的对齐”都会被时间错位掩盖。第二，BOLD 信号存在血氧动力学延迟与时间平滑，因此刺激特征需要在时间上做合适的聚合与延迟建模（例如 FIR 延迟拼接），以使线性模型能够在较低的复杂度下捕捉到主要的响应动力学。

本实验的研究动机与既有工作在问题设置上高度一致。自然语音刺激下的体素级语义地图研究表明，语义信息在皮层上呈现跨网络的分布式组织，并且可以通过线性编码模型在未见刺激上进行验证 [4]；进一步的语义关系映射工作指出，大脑对概念与关系的表示并非局部模块化，而是以重叠与梯度形式分布在多处皮层区域 [12]。在模型侧，大规模集成比较研究显示 Transformer 架构的语言模型在脑与行为数据上的拟合度与其语言建模能力存在稳定联系 [9]，但也存在“对齐来源于特征发现而非预测编码”的替代解释 [1]。因此，本报告在既定数据集与评估框架下，通过对比文本、音频与多模态预训练模型的多层特征，给出一组可复现的对齐结果，并在 ROI 统计与脑图可视化层面补充对区域偏好的观察。

本项目的研究目标是比较不同预训练模型、不同层的特征对于大脑反应的可预测性，并进一步分析不同脑区对不同模态特征的偏好。这里“可预测性”采用编码模型预测值与真实 fMRI 的相关系数作为度量；相关越高，表示该特征在当前编码框架下与脑信号对齐程度越高。本项目的实现强调可追溯性：每一次模型与层的评估都会在 `results/` 下产生对应的 `log.txt` 与 `corr_layer*.npy` 文件，前者记录多被试的均值与标准差，后者保存每个 ROI 的相关图（corr map）。统计图由 `report/scripts/make_figures.py` 直接读取 `results/summary.csv` 与 `results/roi.csv` 生成；脑图由 `src/run_plot_corr_maps.py` 读取 `corr_layer*.npy` 生成，并保存到 `report/figures/brainmaps/`，从而保证“数值结果—可视化—原始文件路径”三者可以互相核验。

需要明确本报告的边界：本报告仅描述当前仓库中已生成并保存为文件的实验结果，不对尚未运行、运行失败、或未保存为结果文件的实验做任何陈述。尤其是非线性编码模型部分，在当前结果目录中未形成可用于对比的系统性输出，因此本报告只讨论线性编码与线性融合（特征拼接）在现有结果上的表现，并在讨论中说明未覆盖部分。

13

数据、对齐表与 TR 级刺激构建

本项目使用的原始文件位于 `data/raw/`。fMRI 数据以 ROI 形式预先整理为 `21styear_all_subs_rois.npy`，对齐表位于 `21styear_align.csv`，音频刺激位于 `21styear_audio.wav`。`src/data.py` 将这些文件加载为可被特征抽取与编码建模直接使用的结构：`load_fMRI()` 返回一个以被试编号为键的字典，每个条目是形状为 $(T, 360)$ 的矩阵，表示 T 个 TR 上 360 个 ROI 的 fMRI 信号；`load_audio()` 以固定采样率读取整段音频；`load_align_df()` 读取对齐表并为每个词构造 TR 编号。

对齐表 `21styear_align.csv` 每行包含四列：保留大小写的词、全部小写的词、词开始时间戳（秒）与词结束时间戳（秒）。对齐表中存在缺失项，代码对时间戳进行向后填充，并将缺失词以 `None` 作为占位。随后根据 TR 时长将词级时间戳映射到离散 TR 索引。项目设置 TR 为 1.5 秒，因此对于词开始时间 t （单位秒），对应 TR 索引为 $[t/TR]$ 。这一步的输出是一个包含 `tr` 列的数据框，它把每个词归入某一个 TR，从而为后续“把词级特征聚合为 TR 级特征”提供了确定的分组键。

TR 级刺激构建需要同时处理三条时间轴：词级时间轴（用于文本与文本端对齐）、连

续波形时间轴（用于音频分片）、TR 采样时间轴（用于与 fMRI 对齐），以及 BOLD 延迟轴（用于 FIR 延迟展开）。本项目对文本与音频采取统一的策略：先在原始粒度上抽取预训练模型特征，再将特征聚合到 TR。对于文本而言，`src/text_pipeline.py` 先为每个词构造上下文窗口（默认 200 token），输入语言模型得到词级或 token 级表征，随后按 `tr` 分组，对同一 TR 中所有词的表征求均值得到 TR 级文本特征。运行时可能出现 pandas 的 FutureWarning，这属于 API 行为变更提示，不影响当前版本下的数值计算与输出文件。

对于音频而言，连续波形根据 TR 窗口切分为一系列 chunk。配置 `src/config.py` 中的 `AUDIO_SR=16000` 表示采样率为 16kHz；当窗口设置为 1TR、2TR、3TR、6TR 时，分别对应 1.5s、3.0s、4.5s、9.0s 的音频片段。每个片段作为一个输入样本送入音频模型得到表示，形成与 TR 一一对应的序列。窗口长度不仅决定了音频表征是否覆盖跨 TR 的韵律与语音单位结构，也会与 FIR 延迟展开共同决定“刺激历史覆盖范围”，因此音频模型部分会系统比较不同 TR 窗口的效果。

为了使后续编码模型稳定训练，特征在进入回归之前会进行降维。当前实现默认使用 PCA 将 TR 级特征降到 250 维 (`DEFAULT_PCA_DIM=250`)，其目的在于减轻高维特征与有限样本量组合导致的病态问题，并降低回归求解成本。所有预处理都在特征与脑信号完成 TR 级对齐之后进行，从而保证特征矩阵与 fMRI 的时间轴严格一致。

14

预训练模型、层选择策略与特征提取实现

14.0.1 模型集合与本次报告覆盖范围

本项目将刺激表示划分为三类：文本模型表示、音频模型表示与多模态模型表示。文本模型部分在当前结果中覆盖 `gpt2`、`bert-base-uncased` 与 `roberta-base`；音频模型部分覆盖 `facebook/wav2vec2-base-960h`、`microsoft/wavlm-base-plus` 与 `facebook/hubert-base-ls960`；多模态模型部分覆盖 `Whisper` (`openai/whisper-small`、`openai/whisper-base`) 与 `CLAP` (`laion/clap-htsat-unfused`)。以上模型的可比较结果体现在 `results/summary.csv` 中，并且每一条统计记录都可以追溯到对应的 `results/.../log.txt` 与 `corr_layer*.npy` 文件。

在方法论上，本项目遵循“以预训练模型的多层隐藏状态作为可解释特征空间”的

通行做法。语言模型方面，ELMo 的深层上下文表示 [7]、BERT 的双向 Transformer 表示 [3] 与 GPT-2 的自回归表示 [8] 构成了常用对照组，用以区分不同训练目标与不同上下文利用方式对脑预测性能的影响。语音模型方面，自监督框架 wav2vec 2.0 [2] 提供了从原始波形到高层语音表征的分层表示，并被用于检验模型对齐是否更接近真实语音加工通路 [6]。在更宏观的模型比较工作中，大规模集成建模强调“层选择”对脑拟合度的关键作用 [9, 5]，同时也提醒我们，模型对齐的来源可能既包含预测目标，也包含更一般的特征发现与迁移能力 [1]。

14.0.2 层选择：按相对深度等比例取样

不同预训练模型的层数并不相同，例如多数 base 级 Transformer 编码器为 12 层，而 Whisper-base 的编码器层数更少。如果直接固定使用某些绝对层号（例如一律抽取第 12 层），会导致在浅层模型中越界，或在深层模型中取样过稀，从而让“层对齐差异”混入“层号不匹配”带来的偏差。本项目的层选择采用等比例的相对策略：对每个模型先从配置中读取总层数 L ，再用等间距取样从 1 到 L 选取若干层，并四舍五入去重得到最终层集合。这一策略的直接结果是：对于 12 层模型，典型层集合为 {1, 4, 6, 9, 12}；对于 6 层模型，则可能得到 {1, 2, 4, 5, 6}。因此，本报告中“layer=k”的含义始终是“该模型结构中的第 k 层”，而不是跨模型共享的绝对语义层级；跨模型比较时，我们把其理解为“从浅到深的相对位置”，并结合每个模型的层数解释其表现。

14.0.3 文本特征：上下文窗口与双层池化

文本特征提取遵循“词级上下文—词级表征—TR 级聚合”的流程。`src/run_text_models.py` 以对齐表为索引，为每个词构造长度为 200 token 的上下文窗口 (`ctx_words=200`)，并将“预分词后的词序列”输入 HuggingFace 模型得到隐藏层输出。代码层面存在两次池化：第一次发生在模型输出端，用于将 token 序列压缩为一个词窗口的向量，支持最后 token 表征或 token 平均；第二次发生在时间对齐阶段，即将同一 TR 内所有词的向量做平均得到 TR 级表示。当前结果文件对应的实现采用“模型端取 last token 表征，TR 内对词向量做平均”的组合，这与自然语言理解中“当前词由其左侧上下文决定”的建模假设一致，并且可以将变长词序列稳定映射到定长向量。

14.0.4 音频特征：TR 窗口切分与帧级池化

音频特征提取遵循“波形分片—模型表征—窗口池化”的流程。整段音频以 16kHz 采样率读取后，按 TR 窗口切分为若干 chunk；每个 chunk 输入音频模型得到时间序列隐藏状态，再用 attention mask 对有效帧做平均池化得到单个向量。当前结果系统比较了 1TR、2TR、3TR、6TR 等多种窗口长度，目的在于检验更长的声学上下文是否有助于在 BOLD 延迟下提高可预测性。

14.0.5 多模态特征：模型内部融合与 TR 对齐

多模态模型的关键区别在于其输出不是“纯音频编码器的表示”，而是模型结构中显式对齐或融合了文本与音频信息后的表示。Whisper 属于编码器—解码器结构，本项目使用其编码器侧的表示作为与输入语音相关的表征来源，并对不同层进行比较；CLAP 同时包含音频与文本编码器，输出位于共享嵌入空间的音频表示，本项目将其视为多模态对齐框架下的表示来源，并在同样的 TR 窗口切分策略下进行评估。由于不同多模态模型对输入形式与采样率存在约束，当前报告仅讨论在 `results/` 中已经成功产出 corr map 的配置。

15

编码模型、评价指标与输出结构

15.0.1 从 TR 特征到 BOLD：线性编码模型的形式化

设某一类特征在 TR 级别上形成矩阵 $X \in \mathbb{R}^{T \times D}$ ，其中 T 为有效 TR 数量、 D 为特征维度；对应被试的 fMRI ROI 信号为 $Y \in \mathbb{R}^{T \times V}$ ，其中 $V = 360$ 为 ROI 数量。线性编码模型采用岭回归，对每个 ROI 同时求解权重矩阵 $W \in \mathbb{R}^{D \times V}$ ：

$$\hat{W} = \arg \min_W \|XW - Y\|_2^2 + \alpha \|W\|_2^2. \quad (1)$$

这里 α 为 L_2 正则强度。岭回归的优势在于当 D 较大且特征存在共线性时，仍能得到数值稳定的解，并在有限样本下缓解过拟合。当前工程中 `DEFAULT_ALPHAS` 提供了若干候选正则强度，但由于 `DEFAULT_KFOLD=1`，实际训练并未进行 K 折交叉验证选择超参，而是在单次训练/测试划分下使用候选列表中的第一个 α 。因此，当前结果可被理解为一套

“固定正则的线性基线”，其主要价值在于为不同特征与不同层提供统一且可追溯的对比基准。

这种“线性编码 + 严格时间对齐 + 在未见刺激上评估”的组合并非偶然，而是与自然叙事语义地图与集成建模工作在评价逻辑上保持一致。语义地图工作通过正则化线性回归在新故事上预测 fMRI，并据此把“可预测性”作为语义表征存在的证据 [4]。大规模集成建模进一步强调统一评估协议的重要性，用以公平比较不同模型与不同任务，从而把差异尽可能归因到表示本身 [9]。与此同时，关于“预测编码是否是对齐来源”的争论提示我们，编码性能的提升既可能来自预测目标，也可能来自更一般的特征发现与结构归纳，因此需要在多模型、多层次与多窗口条件下稳健比较并保持解释克制 [1, 5]。

15.0.2 BOLD 延迟建模：PCA 降维与 FIR 延迟展开

从预训练模型得到的 TR 特征往往维度较高，直接回归会带来计算成本与病态风险。因此，本项目在回归前对 TR 特征做 PCA 降维，默认保留 250 维 (DEFAULT_PCA_DIM=250)。随后，为显式建模 BOLD 延迟与时间扩散，本项目采用 FIR(finite impulse response) 延迟展开：将每个 TR 的特征与若干个过去 TR 的特征按时间顺序拼接，形成扩展特征矩阵。当前默认设置为窗口长度 4、偏移 1 (DEFAULT_FIR_WINDOW=4, DEFAULT_FIR_OFFSET=1)，这意味着在预测某一 TR 的 fMRI 时，模型可以使用从较早 TR 开始、覆盖若干步历史的刺激表示，从而在不引入非线性结构的前提下捕捉响应延迟。

15.0.3 数据划分与多被试汇总

编码模型以每个被试为单位独立训练与评估：对每个被试的 (X, Y) 在时间轴上做截断以去除边界 TR，然后按时间顺序切分为训练段与测试段（当前实现默认测试比例为 0.2）。在测试段上计算预测信号与真实信号的相关系数，得到长度为 360 的相关向量 (corr map)。为得到多被试的总体性能，项目对每个被试的 corr map 求均值作为该被试的总体分数，再对所有被试分数计算均值与标准差并写入 `log.txt`。因此，报告中“平均值 ± 标准差”对应的是跨被试的统计，而不是跨折的统计。

15.0.4 评价指标与 corr map 的含义

评价指标为 Pearson 相关系数。对第 v 个 ROI，设测试段真实信号为 y_v 、预测信号为 \hat{y}_v ，则相关为

$$r_v = \frac{\sum_t (\hat{y}_{v,t} - \bar{\hat{y}}_v)(y_{v,t} - \bar{y}_v)}{\sqrt{\sum_t (\hat{y}_{v,t} - \bar{\hat{y}}_v)^2} \sqrt{\sum_t (y_{v,t} - \bar{y}_v)^2}}. \quad (2)$$

将所有 ROI 的 r_v 组成向量即可得到 corr map。项目保存的 `corr_layer*.npy` 即为该向量，长度为 360，前 180 对应左半球 ROI，后 180 对应右半球 ROI。可视化时，`src/viz.py` 读取 HCP-MMP 的 ROI 标签文件，将 ROI 相关值映射回 `fsaverage` 表面顶点并绘制，输出以左右半球的外侧与内侧视图组成的四联图，保证角度稳定且信息密集。

15.0.5 输出文件结构与可追溯性

本项目输出结构以 `results/` 为根。文本、音频、多模态的线性编码结果分别位于 `results/text/`、`results/audio/`、`results/multimodal/`；每个配置目录包含 `log.txt` 与若干 `corr_layer*.npy`。融合结果位于 `results/fusion/`，其中每个融合对目录包含融合日志与多个融合 corr map 文件（例如 `corr_t9_a6_ctx200_tr1.npy`）。报告中的统计图由 `report/scripts/make_figures.py` 从 `results/summary.csv` 与 `results/roi.csv` 生成，脑图由 `src/run_plot_corr_maps.py` 从 corr map 生成并保存到 `report/figures/brainmaps/`。该设计使得报告中每一张图都能追溯回唯一的源文件路径，便于复核与增量补充实验。

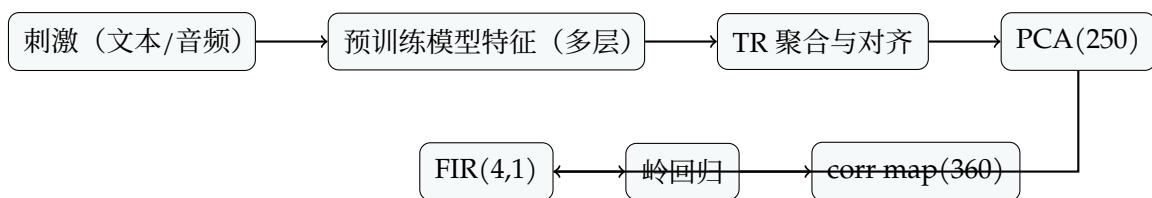


图 1 线性编码建模流水线概览。为避免版面溢出，示意图采用两行布局，但顺序与实现一致。

16

文本模型结果：层次对齐与区域分布

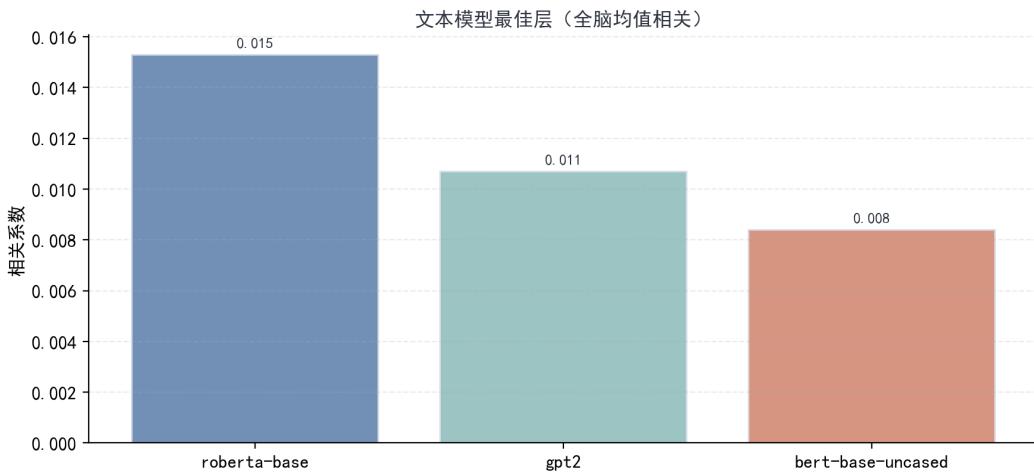


图 2 文本模型最佳层的全脑均值相关系数（由 `report/scripts/make_figures.py` 从 `results/summary.csv` 生成）。

文本模型部分的所有结论均来自 `results/summary.csv` 中 `/text/` 相关条目，以及相应目录下的 `corr_layer*.npy`。本次结果对应的文本上下文窗口固定为 200 token (目录名 `win200`)，词级表示在对齐阶段按 TR 内平均得到 TR 级输入特征。该设置直接决定了文本表示所覆盖的语义时间尺度：当叙事结构跨越更长时间范围时，固定窗口可能截断长程依赖；当 TR 内词数较少时，TR 内平均会引入更强的采样噪声。尽管如此，这一固定设置为跨模型的层比较提供了可复现的对照条件。

图 2 展示了三种文本模型各自最佳层在全脑均值相关上的对比。当前完成的结果显示，`roberta-base` 的最佳层为 layer4，均值相关约为 0.0153；`gpt2` 的最佳层为 layer12，均值相关约为 0.0107；`bert-base-uncased` 的最佳层为 layer6，均值相关约为 0.0084。就数量级而言，文本模型在本任务中的可预测性明显弱于音频与多模态模型。由于本项目没有在同一模型中显式控制“音频线索是否存在”，因此我们在此不对“语义贡献是否被声学驱动掩盖”做超出已完成结果的推断，而是把文本结果视为在当前时间对齐与线性框架下的一组可追溯基线。

为了满足“同一类模型的脑图对照”这一展示要求，图 3 将本次文本模型各自最佳层的 `corr map` 统一绘制并组合成一张对照图。该图使用同一套 ROI 到顶点映射与同一色标策略，能够直观看到文本模型在空间分布上的共同点与差异。作为更细粒度的示例，图 4 给出 RoBERTa 最佳层 (`win200, layer4`) 的单独脑图，图 5 给出相同配置下 ROI 层面的 Top20，用于与后续音频、多模态结果在区域偏好上做直接对照。

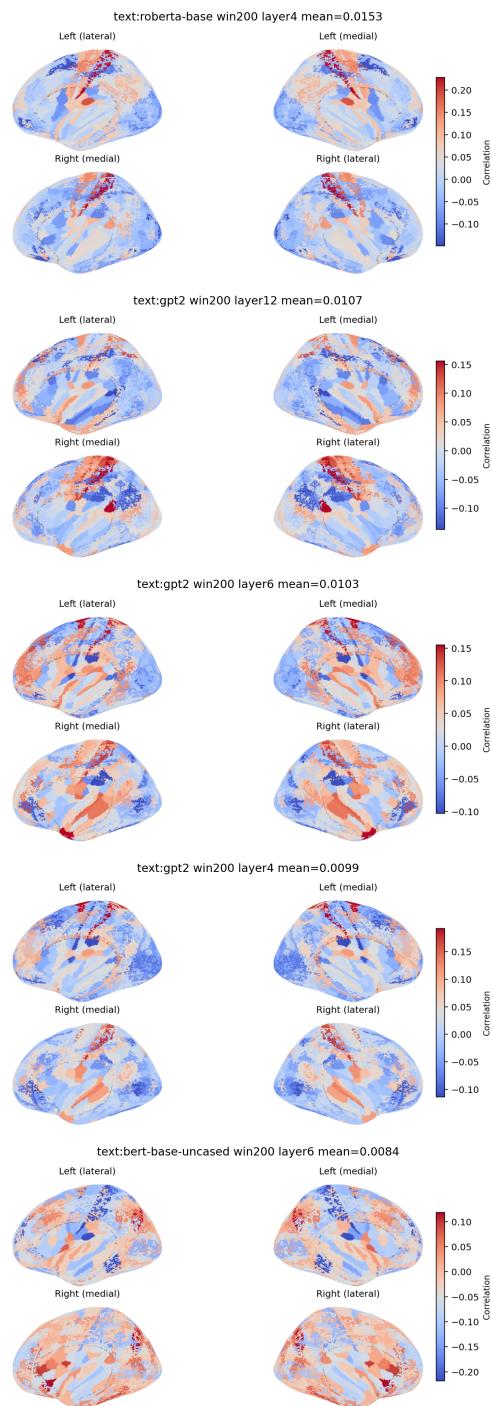


图3 文本模型类别脑图对照：RoBERTa (win200, layer4)、GPT2 (win200, layer12)、BERT (win200, layer6) (由 `src/run_plot_corr_maps.py` 从对应 `corr_layer*.npy` 绘制并组合)。

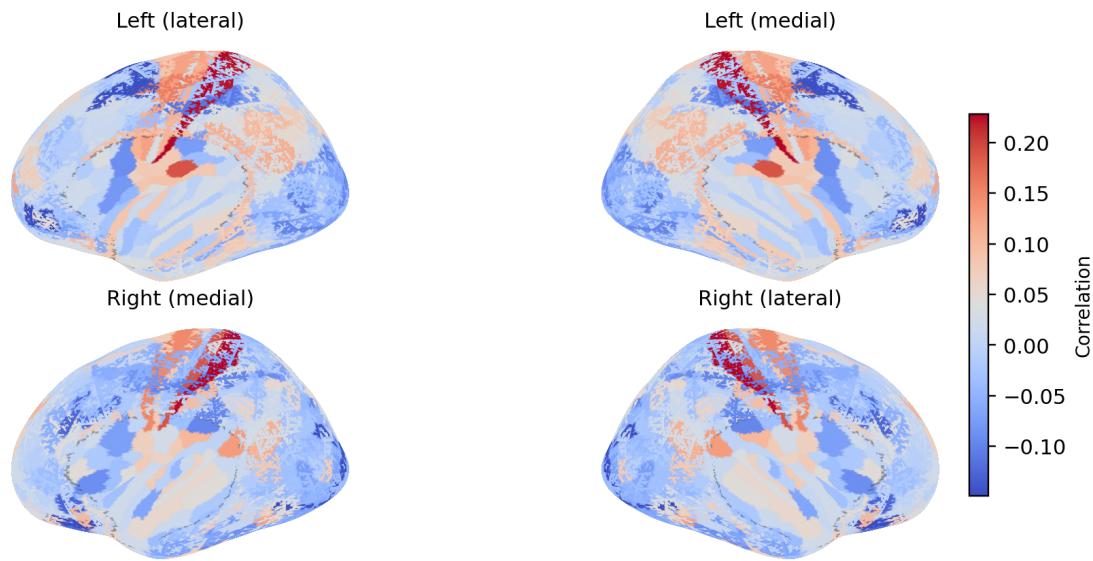


图 4 RoBERTa-base (win200, layer4) 相关图可视化（由 `src/run_plot_corr_maps.py` 从 `corr_layer4.npy` 绘制）。

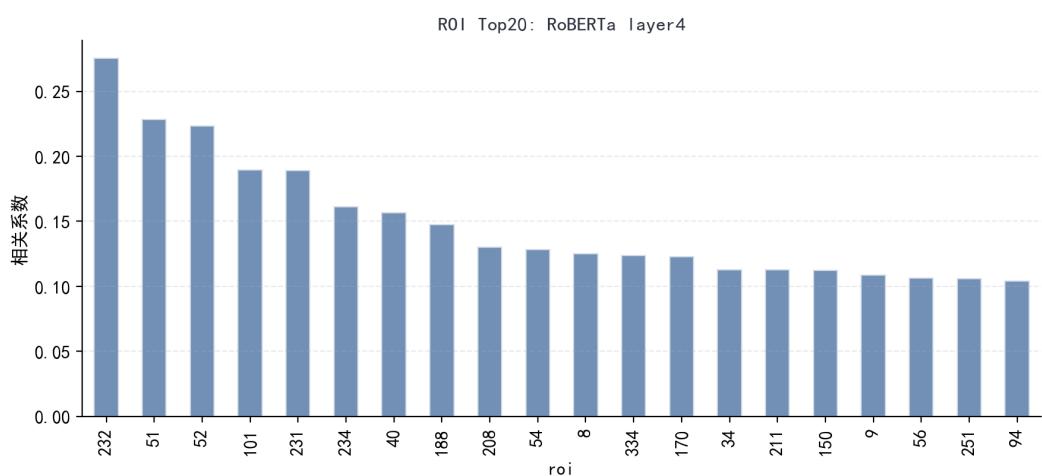


图 5 RoBERTa-base (win200, layer4) 对应 corr map 的 ROI Top20 (由 `report/scripts/make_figures.py` 从 `results/roi.csv` 生成)。

17

音频模型结果：TR 窗口效应与最佳层比较

音频模型部分覆盖三种预训练声学模型，并在 1TR、2TR、3TR、6TR 四种窗口下评估多个层的编码性能。图 6 给出每个音频模型在所有已完成设置中取得的最佳层均值相关。当前结果中，`microsoft/wavlm-base-plus` 在 6TR 条件下的 layer9 达到 0.0916，为音频组内最优；`facebook/wav2vec2-base-960h` 在 6TR、layer9 达到 0.0895；`facebook/hubert-base-ls960` 在 6TR、layer9 达到 0.0823。该排序在数值上与模型家族的预训练目标与结构差异一致，但报告不将其过度解释为结构因果，仅作为实证对比的结论陈述。

为了进一步回答“窗口长度是否系统影响预测性能”，图 7 将每个音频模型在每个 TR 窗口下的最佳层均值相关串联成趋势曲线。对于三种模型，窗口从 1TR 增至 6TR 都带来显著提升：例如 WavLM 在 1TR 的最佳均值约为 0.0316，而在 6TR 上升到 0.0916；Wav2Vec2 在 1TR 的最佳均值约为 0.0274，而在 6TR 上升到 0.0895；HuBERT 在 1TR 的最佳均值约为 0.0326，而在 6TR 上升到 0.0823。该趋势说明长时间窗的声学聚合是当前设置下提升编码性能的关键因素，且这种提升并非某一个模型的偶然现象。

空间层面上，本报告需要同时满足两类展示要求：一类是“同一类别模型的脑图对照”，另一类是“最优模型的高质量脑图”。图 8 将音频类别中三种模型的最佳配置脑图并置，便于观察在相同绘图视角与色标下的分布差异；图 9 则单独展示 WavLM 的最优配置（6TR，layer9），并在图 10 给出 ROI Top20 作为区域偏好分析的入口。由于音频模型在当前结果中整体最强，其空间分布也作为后续多模态与融合分析的基线参照，用于判断多模态表征是否在相同脑区或不同脑区带来额外增益。

18

多模态模型结果：Whisper

多模态模型部分的结果来自 `results/summary.csv` 中 `/multimodal/` 相关条目以及对应目录下的 `corr_layer*.npy`。图 11 展示了已完成的多模态模型在最佳层上的全脑均值相关，图 12 展示了不同 TR 窗口下的最佳层趋势。当前结果中，`openai/whisper-base` 在 6TR、layer2 达到 0.0889，已经非常接近强音频基线

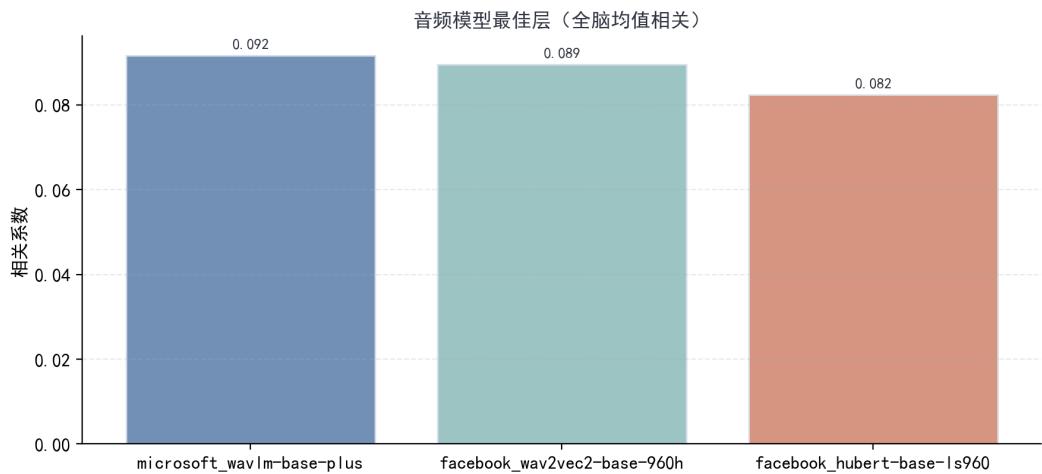


图 6 音频模型最佳层的全脑均值相关系数（由 report/scripts/make_figures.py 从 results/summary.csv 生成）。

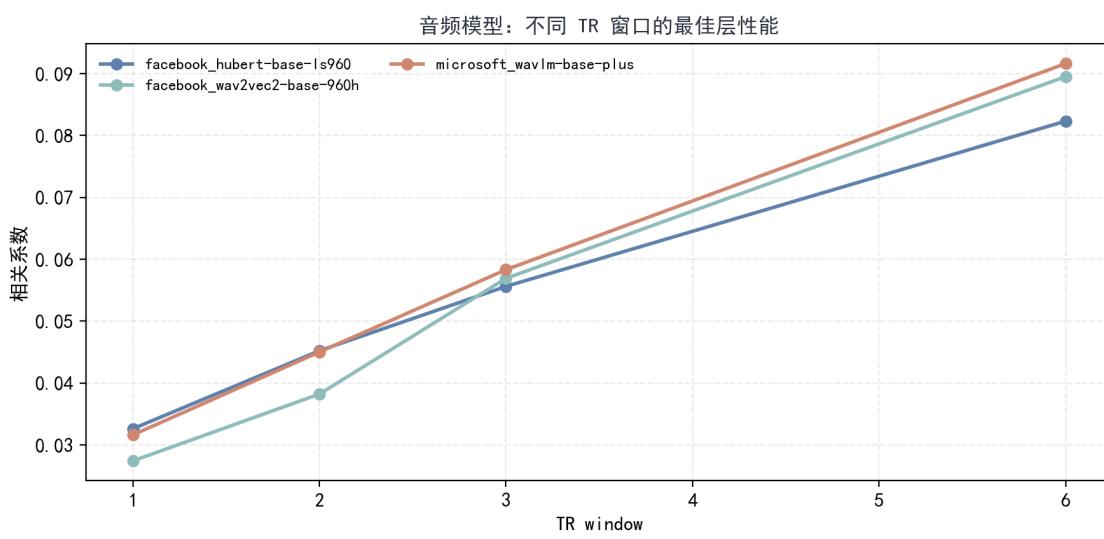


图 7 音频模型在不同 TR 窗口下的最佳层性能趋势（由 report/scripts/make_figures.py 从 results/summary.csv 聚合生成）。

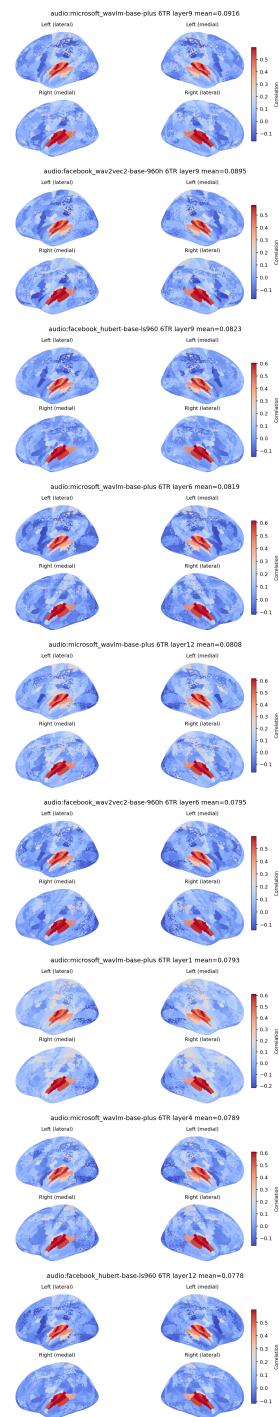


图 8 音频模型类别脑图对照：WavLM、Wav2Vec2、HuBERT 在各自最佳配置下的相关图（由 src/run_plot_corr_maps.py 绘制并组合）。

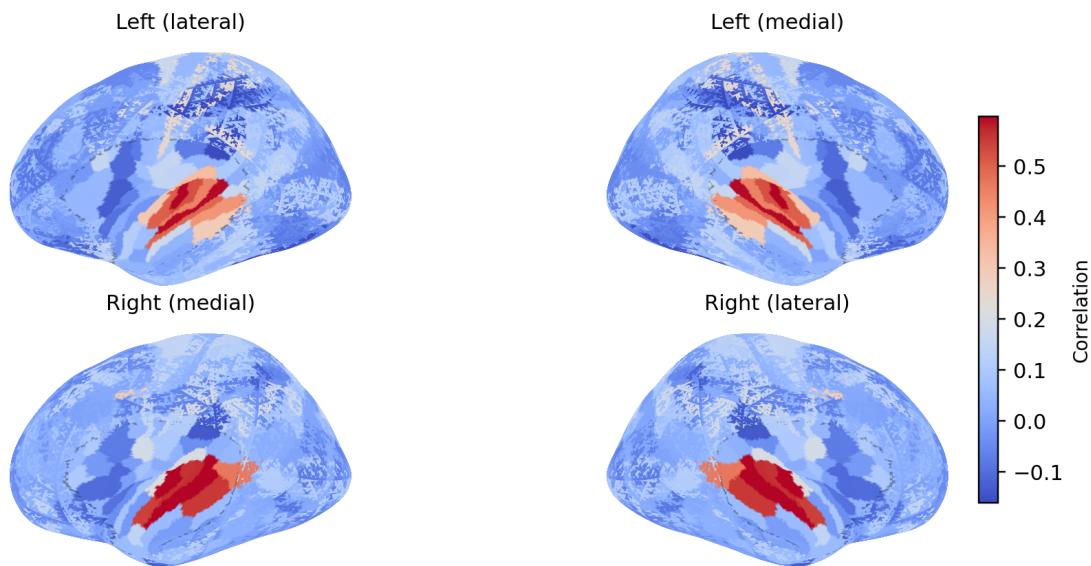


图 9 WavLM-base-plus (6TR, layer9) 相关图可视化（由 `src/run_plot_corr_maps.py` 从 `corr_layer9.npy` 绘制）。

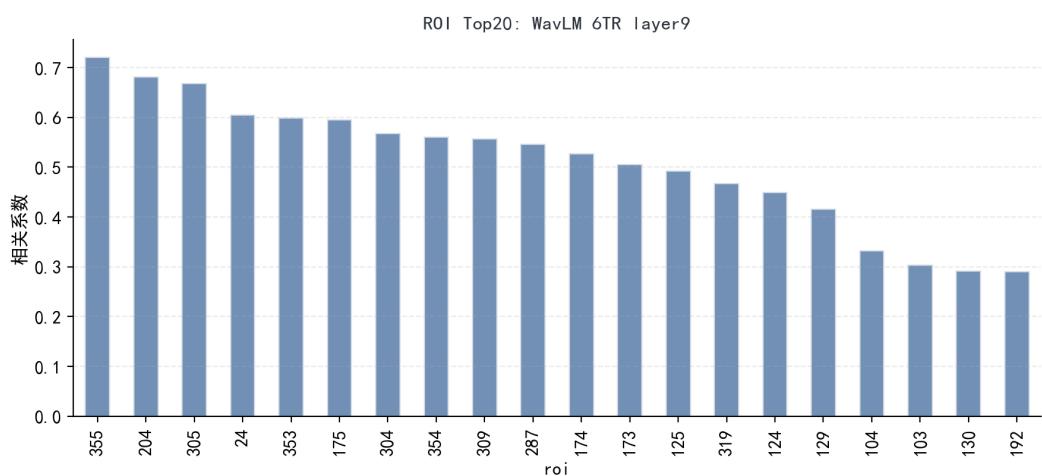


图 10 WavLM-base-plus (6TR, layer9) 对应 corr map 的 ROI Top20 (由 `report/scripts/make_figures.py` 从 `results/roi.csv` 生成)。

Wav2Vec2 的 0.0895; `openai/whisper-small` 在 6TR、layer9 达到 0.0844。值得注意的是，Whisper-base 在 1TR、2TR、3TR 条件下也有多层完成记录，其最佳均值分别为 0.0270、0.0408、0.0579，趋势与音频模型一致，说明多模态模型同样高度依赖较长时间窗的聚合。

在“多模态是否带来额外优势”的讨论中，需要区分两类基准。第一类是纯文本语义空间的编码能力，它通常在自然叙事数据上较弱，且对上下文窗口与层选择敏感 [9, 1]。第二类是纯语音表征空间的编码能力，自监督语音模型在多层表示中形成从声学到更抽象结构的梯度，被认为更接近真实的语音加工链路 [2, 6]。Whisper 作为编码器—解码器框架，其编码器输出仍主要由语音输入决定，因此它是否能超越强音频基线，取决于其内部是否形成了更贴近语义层级的压缩表示。与解码方向的工作相比，Whisper 并不直接生成“语义重构”，但连续语义重构研究表明，fMRI 可以在语义层面约束生成模型输出，这为未来将多模态模型用于解码提供了方法参照 [10]。

从“是否优于音频基线”的角度来看，当前已完成的多模态结果并未显著超越音频最优（WavLM 0.0916），但 Whisper-base 已经达到与 Wav2Vec2 接近的水平。由于本报告不引入未完成的显著性检验或噪声天花板估计，因此在解释上保持克制：可以确认多模态模型在当前设置下具有较强的可预测性，但不能仅凭均值相关就断言其“比音频更语义化”或“更接近高阶语言区”，这些需要结合 ROI 命名映射与进一步统计。

与文本与音频章节一致，本章节同样提供“类别脑图对照”与“最优模型脑图”两类图像。图 13 将 Whisper-base 与 Whisper-small 在各自最佳配置下的 corr map 并置，以便观察不同模型在空间分布上的共性与差异。图 14 展示 Whisper-base 最优配置（6TR, layer2）的单独脑图，图 15 展示其 ROI Top20，用于与音频最优模型的 ROI Top20 做直接比较。由于 `results/roi.csv` 当前只记录 ROI 编号而未提供解剖名称映射，本报告在区域解释上不引入超出文件所能支持的命名推断，而是把重点放在可复现的数值与分布差异描述上。

19

文本 + 音频融合结果：覆盖范围、最优配置与层交互结构

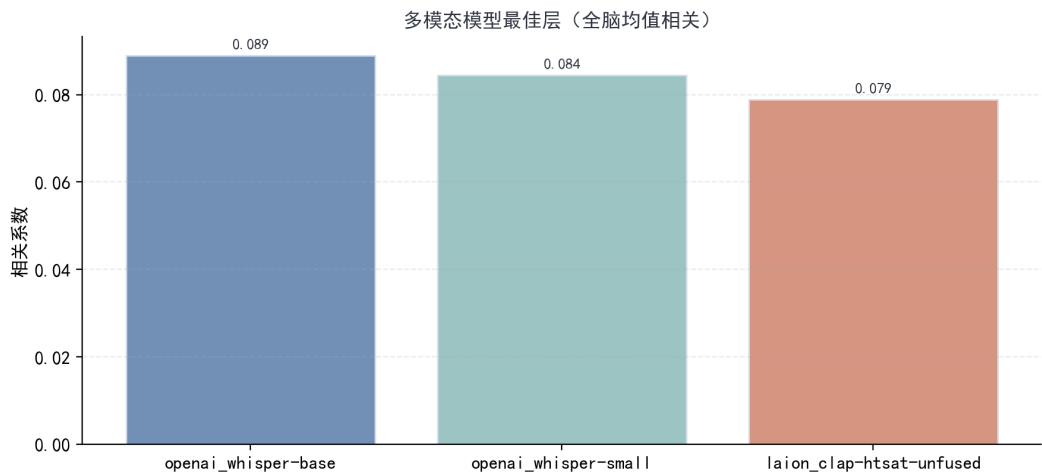


图 11 多模态模型最佳层的全脑均值相关系数（由 report/scripts/make_figures.py 从 results/summary.csv 生成）。

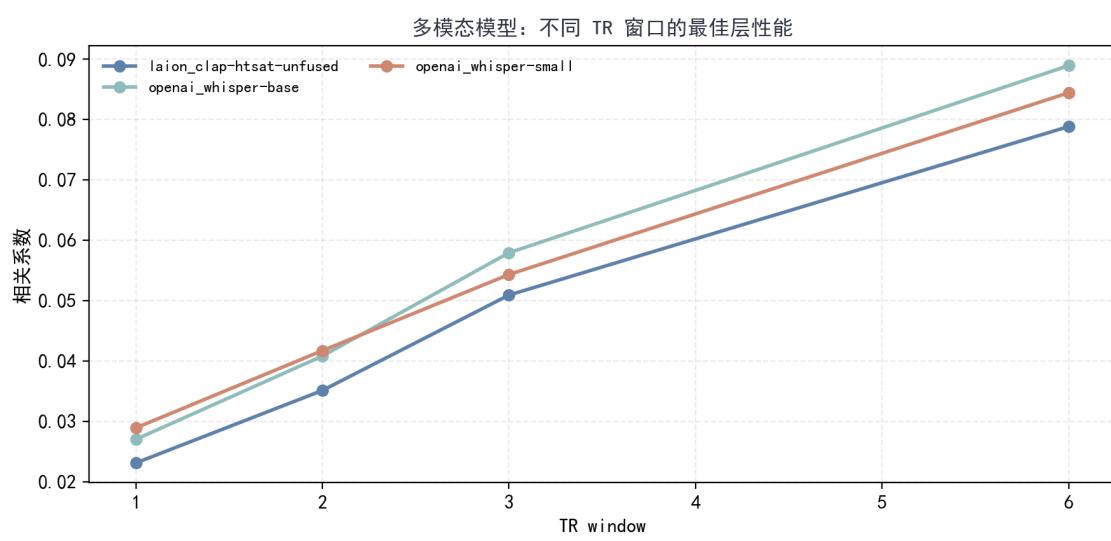


图 12 多模态模型在不同 TR 窗口下的最佳层性能趋势（由 report/scripts/make_figures.py 从 results/summary.csv 聚合生成）。

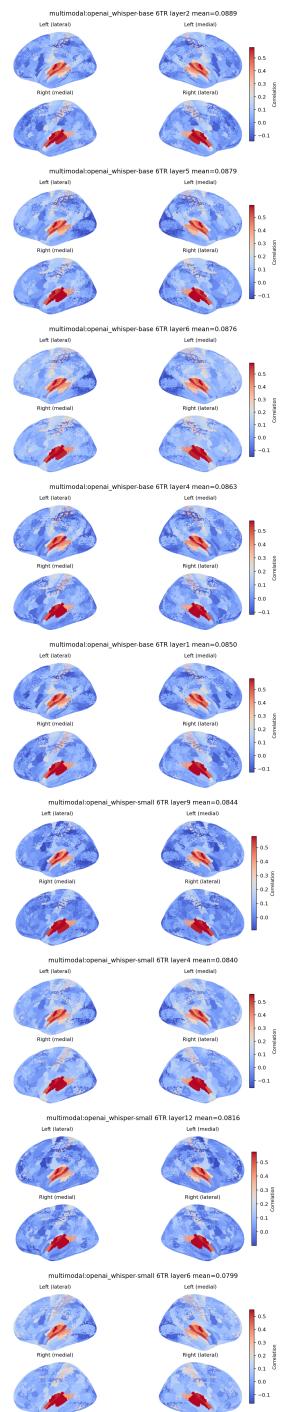


图 13 多模态模型类别脑图对照:Whisper-base(6TR,layer2)与 Whisper-small(6TR,layer9)(由 src/run_plot_corr_maps.py 绘制并组合)。

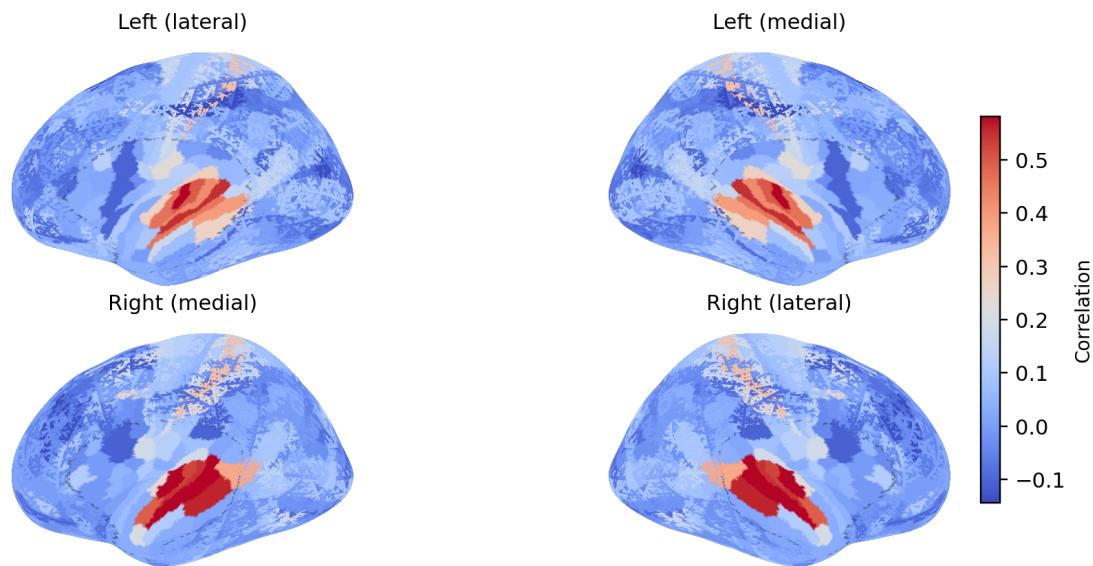


图 14 Whisper-base (6TR, layer2) 相关图可视化 (由 `src/run_plot_corr_maps.py` 从 `corr_layer2.npy` 绘制)。

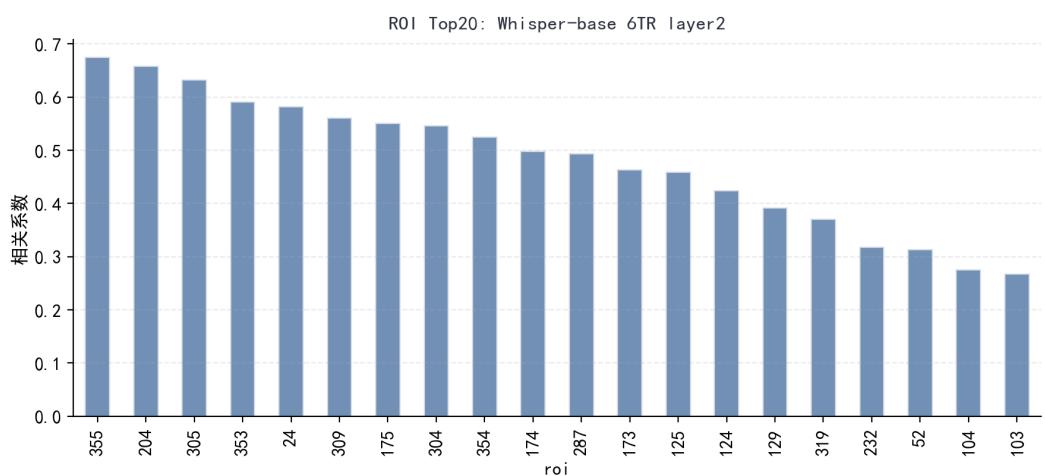


图 15 Whisper-base (6TR, layer2) 对应 corr map 的 ROI Top20 (由 `report/scripts/make_figures.py` 从 `results/roi.csv` 生成)。

融合实验的输出位于 `results/fusion/`。与单模态结果的 `corr_layer*.npy` 命名不同，融合结果的文件名包含文本层、音频层、文本上下文窗口与音频 TR 窗口，例如 `corr_t6_a9_ctx200_tr3.npy`。融合的核心思想是把文本与音频在 TR 级对齐后进行特征拼接，再在同一套 PCA+FIR+ 岭回归框架下评估其对 fMRI 的可预测性。实现上，`src/run_multimodal_fusion.py` 对文本与音频分别做标准化(StandardScaler)，将两者在维度上拼接得到融合特征，再在拼接后的联合空间执行 PCA（默认 250 维），最后做 FIR 延迟展开并回归到 360 个 ROI。由于 PCA 在联合空间上进行，主成分同时反映文本与音频的方差结构，因此“融合是否有效”不仅取决于单模态信息是否互补，也取决于联合空间中哪些方向更易被线性回归利用。

与此前仅有少量融合记录不同，当前目录中融合已经形成较大规模的可追溯输出：在 `ctx=200` 的固定设置下，融合覆盖 `tr_win=1/2/3` 三种窗口、3 个文本模型与 3 个音频模型，以及多层组合，最终在 `results/fusion/` 下生成大量 `corr_t*_a*_ctx*_tr*.npy` 文件并写入对应的 `log.txt`。由于实际可运行的层集合受到“是否存在对应特征文件”的约束，融合并非严格的全笛卡尔积；但在当前结果中，融合日志已经包含 540 条可解析的配置记录，这使得我们可以直接在融合内部比较不同 TR 窗口、不同模型对与不同层组合的影响。

图 16 给出融合中均值相关最高的 Top12 配置，图 17 给出融合在不同 TR 窗口下的全局最优趋势。两张图共同表明：融合最优值随着 TR 窗口增大而显著上升，当前全局最优出现在 `tr_win=3`。基于融合日志的扫描，融合的全局最优配置为 `bert-base-uncased` 与 `microsoft/wavlm-base-plus` 的组合，其均值相关为 0.0535 (标准差 0.0216)，对应 `text_layer=6`、`audio_layer=9`、`ctx_words=200`、`tr_win=3`。该数值仍低于音频与多模态在 6TR 条件下的 0.08–0.09 量级结果，因此融合是否能在更长窗口下进一步接近或超过强音频基线，需要在后续补齐 `tr_win=6` 与更完整特征覆盖后才能回答。本报告在此仅对目前已生成的融合结果进行严格陈述与可视化总结。

融合的重要问题不是“是否只提升一个数值”，而是“文本层与音频层是否存在交互”。为此，图 18 以全局最优模型对为例，在固定 `ctx=200`，`tr=3` 的条件下绘制 `text_layer` 与 `audio_layer` 的二维性能热图。该图在数值上展示了一个稳定事实：不同层组合的性能并非单调随层数增加而提升，而是存在若干局部最优区域，提示融合效果依赖于两种表征的“相对抽象层级匹配”，而非简单地拼接任意深层即可获益。

空间层面上，图 19 展示融合 Top6 配置的脑图对照，图 20 展示全局最优融合配置的单独脑图。两类图像的作用是把融合结果纳入与单模态一致的“统计图—ROI 图—脑图”证据链，使得后续可以在相同视角下对比融合与单模态的空间分布差异，并检查融合是

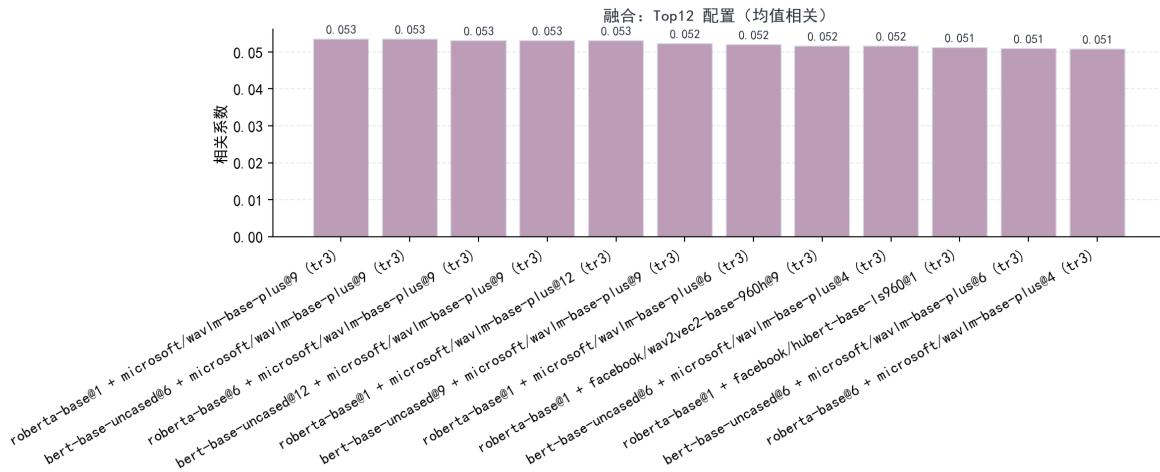


图 16 融合: Top12 配置的全脑均值相关对比 (由 report/scripts/make_figures.py 从 results/fusion/**/log.txt 解析生成)。

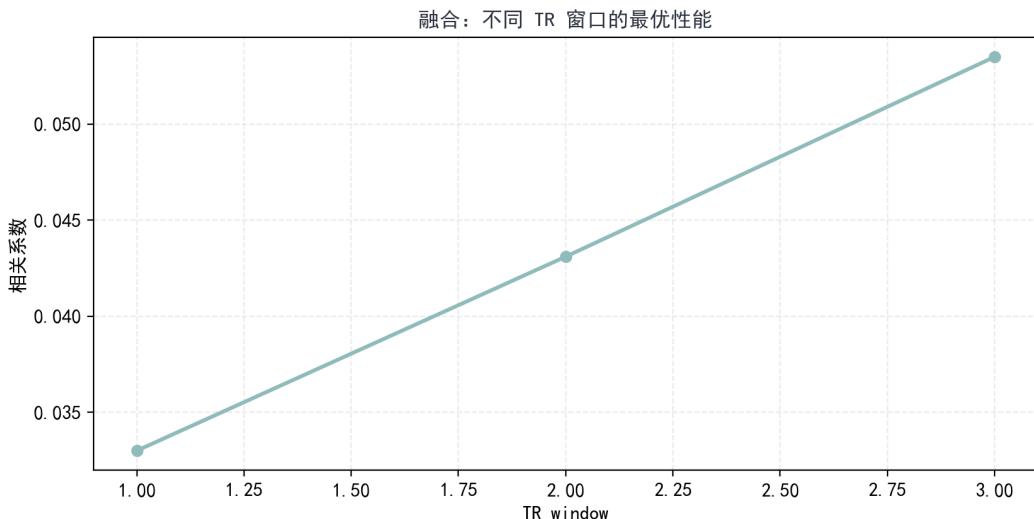


图 17 融合: 不同 TR 窗口下的全局最优性能趋势 (由 report/scripts/make_figures.py 从融合日志解析生成)。

否在部分 ROI 上更接近文本或音频的模式。

20 ROI 分析与综合讨论

ROI 分析的目标是回答“不同特征在皮层不同脑区的预测性能是否存在系统差异”。在当前工程实现中, fMRI 已经以 HCP-MMP 360 ROI 的粒度保存并用于回归, 因此 ROI 分析不是从顶点或体素再聚合到 ROI 的二次统计, 而是对已保存

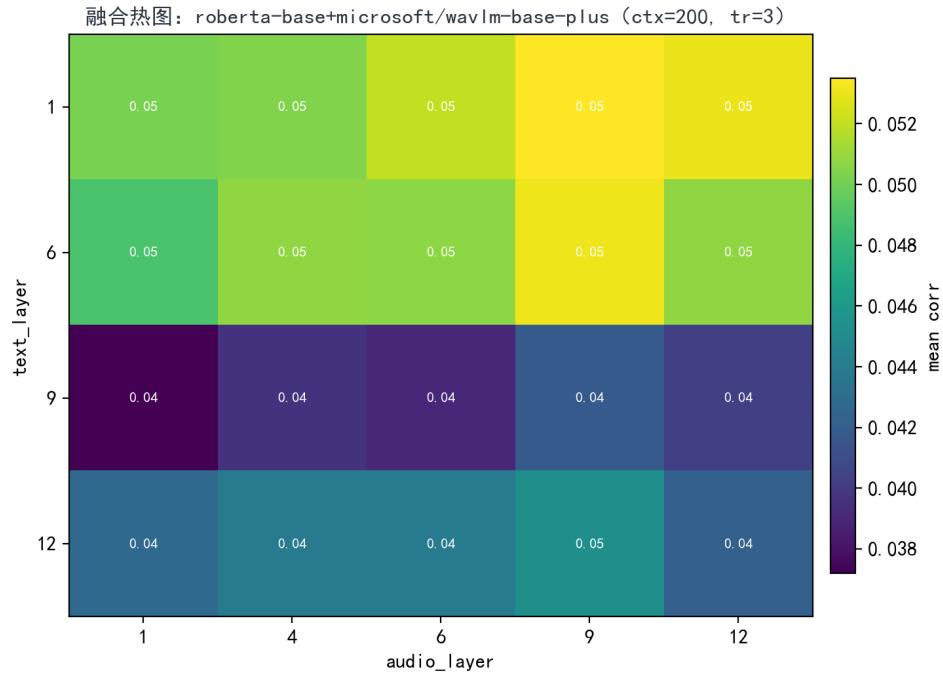


图 18 融合热图：全局最优模型对 (BERT + WavLM) 在 ctx=200、tr=3 条件下的层交互结构 (由 report/scripts/make_figures.py 解析融合日志并绘制)。

的 ROI 相关向量进行整理与排序。具体流程为：`src/run_roi_analysis.py` 扫描 `results/` 下所有 `corr_layer*.npy` 文件，将每个 `corr map` 的 360 个相关值与其源路径一起写入 `results/roi.csv`；随后 `report/scripts/make_figures.py` 读取 `results/roi.csv` 与 `results/summary.csv`，为若干代表性配置绘制 ROI Top20 条形图。由于 `results/roi.csv` 明确记录了每个 ROI 值对应的源文件路径，因此任何一张 ROI Top20 图都可以追溯回唯一的 `corr map` 文件，实现与统计图、脑图一致的可复核链路。

从当前三张 ROI Top20 图的对照可以得到两个在数值层面稳健的观察。第一，当模型整体性能较强（例如音频最优 WavLM 6TR layer9、以及多模态最优 Whisper-base 6TR layer2）时，Top20 ROI 的相关值分布整体上移；当模型整体性能较弱（例如文本 RoBERTa win200 layer4）时，Top20 ROI 的相关也明显较低。该现象与全脑均值趋势一致，提示性能差异并非由单一 ROI 的极端值主导，而更像是多个 ROI 上相关的整体抬升。第二，尽管强模型的 Top20 值整体更高，不同模型的 Top20 ROI 编号与排序并不完全一致，这意味着模型表征的优势可能集中在不同的 ROI 子集上。该差异为“模态偏好与语义偏好”的进一步分析提供了入口，但要把 ROI 编号解释为具体解剖区域或功能系统，需要额外的 ROI 命名映射文件与统计检验（例如多重比较控制）。这些内容在当前结果目录中尚未形成可追溯的输出，因此本报告在此不做超出文件所支持范围的机制性断言。

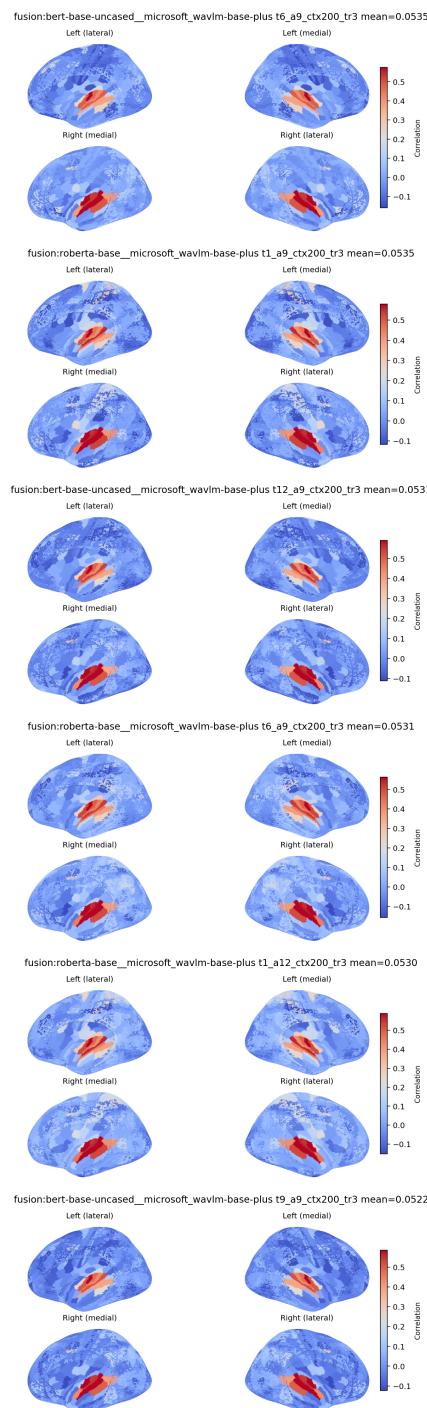


图 19 融合脑图对照：融合 Top6 配置的 corr map 脑图组合（由 src/run_plot_corr_maps.py 从 results/fusion 的 corr_*.npy 选取并绘制）。

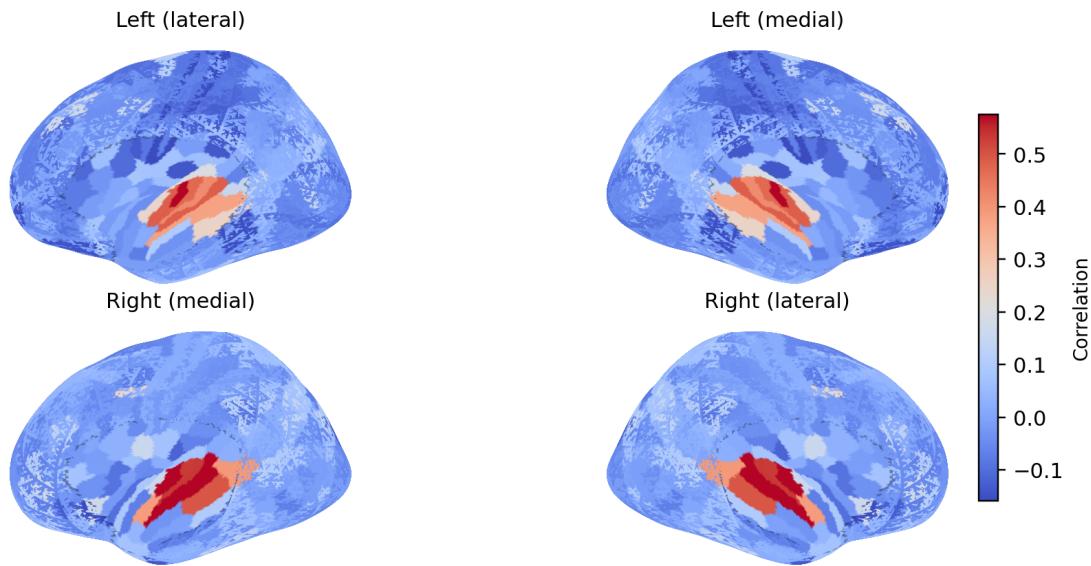


图 20 全局最优融合配置脑图：bert-base-uncased (layer6) + WavLM-base-plus (layer9), ctx=200, tr=3 (由 src/run_plot_corr_maps.py 从对应 corr_*.npy 绘制)。

为了在同一页面中直观对照不同模态的 ROI Top20 分布，本报告将三张代表性 ROI 图并列展示如图 21。该图的阅读方式是先比较数值范围与整体高度，再比较条目 (ROI 编号) 的重合程度，从而形成对“强模型是否带来更广泛的 ROI 提升”以及“不同模态是否偏向不同 ROI 子集”的直接感受。结合前述脑图 (表面空间分布) 与模型对比图 (全脑均值趋势)，ROI 图提供了从空间可视化走向区域定量对照的中间层证据。

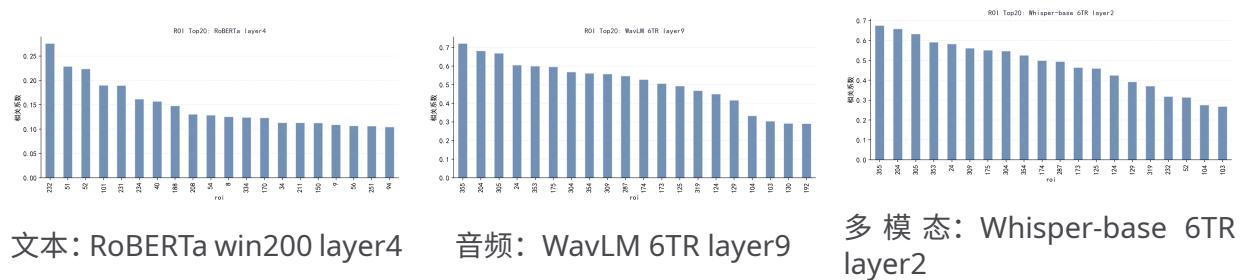


图 21 三种代表性配置的 ROI Top20 对照 (由 report/scripts/make_figures.py 从 results/roi.csv 生成)。

21 结论、局限与后续可扩展方向

在当前已经完成并保存为结果文件的实验范围内，编码模型的可预测性呈现出清晰且可复现的模态差异。文本模型在 win200 的设置下整体相关较低，即便在各自最佳层，其

跨被试均值相关仍处于 10^{-2} 量级；音频模型在较长 TR 窗口下达到约 0.08–0.09 的均值相关，并且不同音频模型在窗口增大时都呈现一致的性能提升趋势；多模态模型中 Whisper-base 在 6TR 条件下取得与强音频基线接近的性能，CLAP 在当前完成范围内略低但同样随窗口增长而提升。上述结论既体现在 results/summary.csv 的均值对比上，也在脑图与 ROI Top20 图上得到一致支持：强模型不仅抬升全脑均值，也使多个 ROI 的相关分布整体上移。

融合实验已经形成大规模可追溯输出。与此前仅能展示少量融合示例不同，当前融合在 `ctx=200` 下覆盖 `tr_win=1/2/3` 三种窗口、3 个文本模型与 3 个音频模型以及多层组合，融合日志可解析记录达到 540 条。融合的全局最优出现在 `tr_win=3`，对应 `bert-base-uncased + microsoft/wavlm-base-plus` 的层组合 (`t6_a9_ctx200_tr3`)，跨被试均值相关达到 0.0535。该数值显著高于文本单模态并在融合内部呈现清晰的窗口效应与层交互结构，但仍低于音频与多模态在 6TR 条件下的 0.08–0.09 量级结果。因此，对“融合是否优于强音频基线”的最终回答仍取决于是否能在更长窗口（例如 6TR）与更一致的特征覆盖条件下完成对等比较。

本报告的第二个重要结论不是某个单一数值，而是本项目的可追溯链路已经被建立并可稳定复用。每一条结果都能从报告中的统计图或脑图回溯到唯一的源文件路径，进一步回溯到生成该文件的脚本与配置，从而使后续扩展实验（例如增加更多文本模型、补齐多模态模型、系统搜索窗口与池化策略、或加入显著性检验）可以在不破坏现有结构的前提下增量进行。与此同时，本报告也明确存在当前结果目录所决定的局限：其一，回归超参与划分策略在当前配置下未进行交叉验证选择，因此结果更适合用于特征对比的基线，而不适合用于对绝对性能做过度外推；其二，ROI 编号尚未映射到解剖名称，限制了对“语义偏好性”的命名解释；其三，非线性编码模型在当前结果目录中未形成与线性结果同结构的 corr map 输出，因此无法纳入同一套图像证据链进行比较。

在上述边界内，本报告已经完成了对现有结果的系统整理：对多模型、多层、多窗口的线性编码结果给出数值对比；对文本、音频、多模态三个类别分别提供“类别脑图对照”与“最优配置脑图”；在 ROI 层面展示代表性模型的 Top20 分布并讨论其可解释性与局限。后续工作的关键并不是对文字叙述做任何“补写”，而是在现有流水线中补齐缺失的实验维度，使这些章节能够在同一模板下自然扩展并与综述合并。

A

附录：结果文件位置与复现说明

本报告的所有数值与脑图均来自项目根目录下的 `results/` 与 `report/figures/`。其中，模型评估与汇总表位于 `results/summary.csv`; ROI 统计位于 `results/roi.csv`; 各模型的相关图数组为 `corr_layer*.npy` (融合为 `corr_*.npy`)，保存在对应的 `results/text/`、`results/audio/`、`results/multimodal/` 与 `results/fusion/` 目录下。统计图由 `report/scripts/make_figures.py` 读取上述结果自动生成到 `report/figures/`; 脑图由 `src/run_plot_corr_maps.py` 读取相关图数组并输出到 `report/figures/brainmaps/`。

参考文献

- [1] Richard Antonello and Alexander G. Huth. Predictive coding or just feature discovery? an alternative account of why language models fit brain data. *Neurobiology of Language*, 2023.
- [2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, 2020.
- [3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.
- [4] Alexander G. Huth, Wendy A. De Heer, Thomas L. Griffiths, Frédéric E. Theunissen, and Jack L. Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458, 2016.
- [5] Sreejan Kumar, Theodore R. Sumers, Tamar Yamakoshi, Ariel Goldstein, Uri Hasson, Kenneth A. Norman, Thomas L. Griffiths, Robert D. Hawkins, and Samuel A. Nastase. Shared functional specialization in transformer-based language models and the human brain. *Nature Communications*, 2024.
- [6] Juliette Millet, Charlotte Caucheteux, Yves Boubenec, Alexandre Gramfort, Ewan Dunbar, Christophe Pallier, and Jean-Rémi King. Toward a realistic model of speech processing in the brain with self-supervised learning. *Advances in Neural Information Processing Systems*, 2023.
- [7] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018.
- [8] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. OpenAI Technical Report, 2019.

- [9] Martin Schrimpf, Idan A. Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45), 2021.
- [10] Jerry Tang, Alexandre LeBel, Shailee Jain, and Alexander G. Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26:858–866, 2023.
- [11] Mariya Toneva and Leila Wehbe. Combining computational controls with natural text reveals aspects of meaning composition. *Nature Computational Science*, 2:745–757, 2022.
- [12] Yanchao Zhang, Alec Tetreault, Yaling Xu, John A. Pyles, and Michael J. Tarr. Connecting concepts in the brain by mapping cortical representations of semantic relations. *Nature Communications*, 11:1–13, 2020.