

认知神经科学

自然故事听觉范式下的语义编码

预训练文本、语音与多模态表征的比较与融合（基于 The
21st Year fMRI 数据）

*Semantic encoding in natural story listening: comparing and fusing pretrained text,
speech, and multimodal representations*

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 24 日

摘要

本报告在自然故事听觉范式的 fMRI 数据上，系统比较了不同预训练模型表征与大脑反应的对齐程度。理论背景部分围绕自然语音语义地图 [4, 12]、集成建模与预测加工框架 [9, 1, 5]、自监督语音表征与皮层层级对齐 [2, 6]、上下文表征与组合语义 [7, 3, 8, 11] 以及连续语义重构的解码路径 [10] 展开。实验部分严格基于当前仓库中已经生成的结果文件撰写，全部数值可追溯到 `results/summary.csv`、`results/roi.csv`、`results/fusion/**/log.txt` 与对应的相关图数组（单模态为 `corr_layer*.npy`，融合为 `corr_*.npy`）。在统一的对齐、降维（PCA）与时延展开（FIR）流程下，我们在多被试上训练线性岭回归编码模型，并以测试集相关系数刻画“可预测性”。结果表明，在当前设置下，音频与多模态模型显著优于纯文本语义表征，并且 TR 窗口长度对音频与融合性能具有决定性影响；在融合实验中，最佳文本层与最佳音频层的组合呈现非单调交互。报告同时呈现统计图与脑图两类证据：统计图由 `report/scripts/make_figures.py` 直接从结果表生成，脑图由 `src/run_plot_corr_maps.py` 将相关图映射到皮层表面生成并保存到 `report/figures/brainmaps/`。



目录

背景与相关工作	3
1.1 背景与研究问题	3
实验设计与方法	6
2.1 实验材料、对齐表与 TR 级刺激构建	6
2.2 模型选择、层抽样与特征提取流程	7
2.3 编码模型、评价指标与输出结构	8
2.3.1 输入特征的标准化与降维	9
2.3.2 FIR 时延展开与线性岭回归	9
2.3.3 训练/测试划分、多被试汇总与指标定义	9
2.3.4 输出文件与可追溯性	10
实验结果	11
3.1 实验结果（一）：文本模型	11
3.2 实验结果（二）：音频模型	11
3.3 实验结果（三）：多模态模型	16
3.4 实验结果（四）：文本-音频特征融合	23
讨论	28
4.1 讨论：结果的意义、不足与展望	28
结论	32
5.1 结论	32

1

背景与相关工作

1.1 背景与研究问题

小组成员与分工

潘宇轩

学号：2023K8009991004

邮箱：panyuxuan231@mails.uca
s.ac.cn

负责整体方案设计与实现；完成文本/音频/多模态与融合特征提取、线性编码模型与评估、ROI 分析与可视化；整理结果并撰写报告。

林俊杰

学号：N/A

邮箱：N/A

未参与本次实验执行与代码实现。

王煜

学号：N/A

邮箱：N/A

未参与本次实验执行与代码实现。

李向阳

学号：N/A

邮箱：N/A

未参与本次实验执行与代码实现。

朱奕

学号：N/A

邮箱：N/A

未参与本次实验执行与代码实现。

本项目研究的问题是：在自然故事听觉范式下，预训练模型得到的刺激表征在多大程度上能够预测全脑 fMRI 响应，并且这种“对齐程度”在不同模态（文本、音频、多模

态)与不同模型层级之间如何变化。该问题的出发点来自两条相互推进的研究线索。第一条线索是自然语音刺激下的语义系统映射。Huth 等使用长时自然故事并构建体素级编码模型,证明在严格的未见故事评估下,线性回归可以得到稳定可复现的语义地图,从而将“可预测性”作为语义表征的可测证据[4]。Zhang 等进一步把分析单位从语义类别扩展到语义关系,指出概念与关系并非由解剖上隔离的模块分别承载,而是通过跨网络的重叠模式编码并支持语义推理[12]。第二条线索来自预训练模型与大脑对齐的集成建模。Schrimpf 等在统一评估协议下系统比较不同架构与不同层,发现 Transformer 模型通常更能解释语言相关脑区的活动,并强调跨模型的可比性与评估一致性[9]。然而,“语言模型拟合脑数据”并不自动推出“大脑在执行下一词预测”。Antonello 与 Huth 通过多项分析指出,模型对齐可以由更一般的特征发现与语言结构归纳解释,且模型内部最擅长预测未来词的层并不必然是最佳脑编码层[1]。因此,本项目把“多模型、多层次”作为基本比较单位,以避免仅凭单一指标外推机制结论。

自然故事听觉范式的另一个关键维度是时间尺度。BOLD 信号相对刺激存在血氧动力学延迟与时间平滑,因此对齐必须同时解决“刺激与 TR 的时间对应”和“刺激对 BOLD 的时延影响”。语义地图研究通常通过把词级语义特征聚合到 TR,并在回归输入端引入时延展开来吸收动力学差异[4]。在本项目中,我们把这一逻辑推广到预训练模型特征:文本侧以 token 上下文窗口构造词级表示,再聚合到 TR;音频侧以 TR 窗口切分语音波形,把短时帧级隐藏状态池化为每个 TR 的表征,再进入同一回归框架。由于时间尺度会影响模型能否捕获长程语境,本项目还系统比较不同 TR 窗口长度对音频与多模态表征的影响,并在融合实验中检验跨模态信息是否互补。

在模型选择上,本项目覆盖三类表征。文本模型选用 GPT-2、BERT 与 RoBERTa,分别代表自回归 Transformer、双向 Transformer 与改进的掩码建模框架;这些模型作为上下文表征的代表,在以往语言脑对齐研究中常被用于区分不同预训练目标对表征结构的影响[7, 3, 8]。音频模型选用 wav2vec2、WavLM 与 HuBERT,它们属于以波形为输入的自监督/弱监督语音表征家族,其中 wav2vec 2.0 的掩码预测目标为“从上下文恢复被遮蔽的离散语音单元”提供了具体实现[2];相关研究表明这类模型在层级上呈现从声学到更抽象结构的渐变,并能在一定程度上对齐皮层的语音处理层级[6]。多模态模型部分覆盖 Whisper 与 CLAP 等模型的可用输出,用于检验“多模态训练或共享嵌入空间”是否能带来超越强音频基线的可预测性,并观察其空间分布是否更接近语义系统。

在研究问题的表述上,本文不把对齐结果直接解释为某一种认知机制的证据,而是以可复现的证据链回答三个可检验的问题。第一,在统一的编码评估框架下,不同模态与不同模型的对齐强弱排序如何,哪些配置构成强基线。第二,在模型内部的层级结构

上，最佳层是否稳定出现在中间层或深层，以及这种层级位置是否与时间窗口长度共同作用。第三，当文本与音频特征进行简单拼接融合时，性能是否出现稳定提升，并且最佳文本层与最佳音频层是否呈现非单调交互。讨论部分将在这些结果约束下结合组合语义与 supra-word 表征观点 [11]、Transformer 功能分化分析 [5]、以及语义重构方向的互补视角 [10]，对结果的意义、不足与后续扩展方向作出解释。

2

实验设计与方法

2.1 实验材料、对齐表与 TR 级刺激构建

本项目的数据由三部分组成：被试在听自然故事时的全脑 fMRI 信号、对应的音频刺激以及文本转写与时间对齐信息。fMRI 的采样以 TR 为单位，项目配置中 TR 时长为 1.5 秒。为了让编码模型的输入与输出处在同一时间轴上，刺激特征必须被构造为长度等于 TR 序列的时间序列，并且每个时间点的特征仅由其对应时间范围内的刺激确定。自然故事范式之所以对对齐提出更高要求，是因为词与声学帧在连续时间中密集出现，且叙事结构会跨越多个 TR 累积；若对齐偏差达到 TR 量级，编码性能会被系统性压低，从而无法解释模型与脑之间的真实差异 [4]。

文本侧对齐采用“词级时间戳 → TR 索引”的映射。对齐表记录每个词在音频中的出现顺序以及其对应 TR 编号，并提供规范化后的词形字段用于构造上下文窗口。本文在特征提取阶段先为每个词构造一个上下文序列，再从预训练语言模型提取词级表示；随后在 TR 对齐阶段，把属于同一 TR 的词向量做平均得到 TR 级文本特征。由于自然故事中不同 TR 的词数并不均匀，某些 TR 可能只有少量词甚至缺失词条，代码实现使用前向填充将缺失 TR 的特征延续为最近的已观测特征，从而得到与 fMRI 序列严格等长的输入矩阵。该处理并不引入新的语义信息，而是把“当前语境”视为在短时间内保持不变的近似，这一近似也与 BOLD 的时间平滑性质一致。

音频侧对齐采用“TR 窗口切分”。首先将整段音频以 16 kHz 采样率读入，随后按 TR 秒数计算每个 TR 对应的采样点数，并以步长为 1 个 TR 的滑动方式切分为长度为 w 个 TR 的音频片段，其中 w 为 TR 窗口长度（例如 1TR、2TR、3TR、6TR）。每个音频片段输入预训练音频模型后得到帧级隐藏状态，再在时间维上做池化得到片段级向量；由于切分步长等于 1TR，片段序列天然与 TR 序列对齐，得到长度为 TR 数的音频特征矩阵。通过改变窗口长度 w ，我们可以直接检验“更长的声学上下文”是否在当前编码框架下提高可预测性，从而把时间尺度作为与层级结构并列的可比较维度。

多模态模型在本项目中以“对齐到 TR 的可用输出”为基本原则：对于 Whisper 类编码器—解码器模型，我们以音频窗口与 TR 对齐的方式得到输入片段，并从模型内部可

获得的隐藏状态构造片段级表征；对于 CLAP 类双编码器模型，我们以相同的音频片段作为输入，并结合对齐表构造同一窗口内的文本字符串，从而在共享嵌入空间中得到可回归的 TR 级多模态表征。无论哪一种模型，最终进入编码模型的输入都被表示为形状为 (T, D) 的矩阵，其中 T 为 TR 数， D 为经池化与降维后的特征维度。

2.2 模型选择、层抽样与特征提取流程

本项目把刺激表征划分为三类：文本表征、音频表征与多模态表征。三类表征共享同一条评估主线：在 TR 对齐后形成 (T, D) 的特征矩阵，经标准化与降维后引入时延展开，再用线性岭回归预测每个 ROI 的 fMRI。该框架的核心假设是：若某个表征捕获了与大脑语言加工相关的信息，则在控制对齐与动力学建模后，它应当在未见刺激上提供更高的预测相关。由于对齐性能对模型层级敏感，本文在模型内部以“多层取样”而非单层取值进行比较，这一做法与集成建模研究对层级结构的重要性强调一致 [9, 5]。

文本模型覆盖 gpt2、bert-base-uncased 与 roberta-base。三者代表了不同的上下文建模机制：BERT 以掩码语言建模学习双向上下文表示 [3]，RoBERTa 在更大规模训练与训练策略上对 BERT 做了系统改进，常被视为更强的双向基线；GPT-2 以自回归方式建模，隐藏状态更直接反映“左侧上下文对当前词的条件化”[8]。在更早的上下文表征研究中，ELMo 通过双向语言模型得到深层上下文词表示 [7]，其思想与本文“以预训练模型中间层作为可解释特征空间”的做法一致。本文实验实际运行的文本模型以 BERT/RoBERTa/GPT-2 为主，ELMo 作为相关工作参照用于解释上下文表征在神经对齐中的位置。

音频模型覆盖 `facebook/wav2vec2-base-960h`、`microsoft/wavlm-base-plus` 与 `facebook/hubert-base-ls960`。这些模型以波形为输入并产生帧级隐藏状态，其共同目标是学习可迁移的语音表征。wav2vec 2.0 的自监督目标通过遮蔽部分时间步并从上下文恢复其离散表示实现 [2]；相关脑对齐研究表明，自监督语音模型在层级上呈现从声学到更抽象结构的渐变，并在一定程度上与皮层语音加工层级对齐 [6]。WavLM 与 HuBERT 在预训练目标与数据增强策略上与 wav2vec 家族有所差异，因而在“时间窗口依赖”与“最佳层位置”上可能呈现不同表现，这正是本文在统一评估协议下比较它们的动机之一。

多模态模型覆盖 `openai/whisper-small`、`openai/whisper-base` 与 `laion/clap-htsat-unfused`。Whisper 属于编码器—解码器结构，编码器把音频转换为序列表示，解码器在文本条件下生成输出；在本文的编码评估中，我们把其内部可

获得的隐藏状态池化为 TR 级向量并进入同一回归框架。CLAP 属于双编码器结构，音频编码器与文本编码器被训练到共享嵌入空间；本文在固定 TR 窗口内构造音频片段与文本片段，以此得到多模态对齐表征。需要强调的是，多模态模型在本项目中的目标不是执行生成任务或检索任务，而是把其内部表征视为一种可比较特征空间；因此，我们只报告在当前实现中已经成功生成并保存 corr map 的配置，不对未完成或未保存的设置做推断。

层抽样策略采用“按相对深度等比例取样”。不同模型的层数并不相同，例如多数 base 级 Transformer 编码器为 12 层，但 Whisper-base 的可用层数更少，CLAP 的可用隐藏状态暴露方式也与标准 Transformer 不同。若直接固定绝对层号，会导致浅层模型越界或跨模型取样不均。本文在每个模型上先读取其总层数 L ，再在 $[1, L]$ 上等间距取样得到若干层，并四舍五入去重，形成该模型的比较层集合。对 12 层模型，该策略通常得到 $\{1, 4, 6, 9, 12\}$ ；对 6 层模型得到 $\{1, 2, 4, 5, 6\}$ 。因此，本文中的“layer= k ”应被理解为“该模型结构内的第 k 层”，跨模型比较时把它视为从浅到深的相对位置，而不是跨模型共享的绝对语义层级。

特征提取在文本与音频侧分别包含两次聚合。文本侧首先构造 token 上下文窗口。本文固定上下文窗口为 200 token：对齐表给出每个词的顺序，代码将当前词之前最近的 200 个 token 作为窗口输入模型。随后在模型输出端进行 token 级池化以得到词窗口向量，本文在实际运行的配置中对自回归模型采用“最后 token 表征”，对双向模型同样使用窗口末端位置对应的表征作为稳定汇聚方式。第二次聚合发生在 TR 对齐阶段：同一 TR 内所有词向量取平均得到 TR 级文本特征。音频侧首先把音频按 TR 窗口切分为片段，模型输出的帧级隐藏状态在时间维上用 attention mask 做均值池化得到片段向量，从而与 TR 序列一一对应。多模态模型的输出同样被池化为 TR 级向量，并在后续步骤与单模态特征共享同一条处理与评估链路。

2.3 编码模型、评价指标与输出结构

本文采用线性编码模型把刺激特征映射到 fMRI 响应。编码建模的选择不是因为线性模型能够穷尽语言加工的非线性机制，而是在自然故事范式下，线性正则化回归提供了一条可复现、可解释且便于跨模型比较的基线。语义地图研究以体素级（或 ROI 级）正则化线性回归在未见故事上预测 fMRI，并据此绘制语义选择性地图 [4]；集成建模工作进一步强调统一评估协议的重要性，使不同模型与不同层的差异更可能来自表征本身而非评估细节 [9]。在机制解释层面，Antonello 与 Huth 的讨论提示我们应避免把编

码性能直接等同于某一种训练目标或单一认知机制，因此本文将编码性能视为一种“表征可预测性”指标，并通过多模型、多层次与多窗口的系统比较减少偶然性 [1]。

2.3.1 输入特征的标准化与降维

对每一种特征，我们先对每一维做标准化，使其在样本维上的均值为零、方差为一。标准化的作用是避免不同特征维度的尺度差异影响岭回归的惩罚项，从而使正则化主要反映信息量而非尺度。随后，我们在特征维上执行主成分分析（PCA），默认降到 250 维。降维的动机有二：其一，预训练模型的隐藏维度通常较高且存在共线性，PCA 有助于稳定回归解；其二，后续时延展开会把特征维度按窗口倍数放大，若不降维会显著增加计算量并降低可重复性。融合实验在拼接后的联合特征空间上执行 PCA，从而使主成分同时反映文本与音频方差结构。

2.3.2 FIR 时延展开与线性岭回归

为处理 BOLD 信号相对刺激的延迟与时间平滑，我们对 TR 级特征做 FIR (finite impulse response) 时延展开。设原始特征矩阵为 $X \in \mathbb{R}^{T \times D}$ ，我们构造延迟窗口长度为 W 、偏移为 O 的拼接特征

$$\tilde{X}_t = [X_{t-O}, X_{t-O-1}, \dots, X_{t-O-(W-1)}] \in \mathbb{R}^{WD},$$

从而把“当前 BOLD”表示为过去若干个 TR 内刺激特征的线性组合。本文在已生成结果对应的默认设置中采用 $W = 4$ 、 $O = 1$ 。在此基础上，我们对每个 ROI 分别拟合岭回归：

$$\hat{Y} = \tilde{X}\beta, \quad \beta = \arg \min_{\beta} \|Y - \tilde{X}\beta\|_2^2 + \alpha\|\beta\|_2^2,$$

其中 $Y \in \mathbb{R}^{T \times R}$ 为 ROI 级 fMRI 响应， R 为 ROI 数。岭回归在高维共线特征下稳定且易于比较，是自然故事编码建模的常用选择。

2.3.3 训练/测试划分、多被试汇总与指标定义

本文的评估以“未见数据上的相关系数”为核心指标。对每个被试，我们先在时间轴两端各排除 10 个 TR，以降低边界效应。随后采用单次训练/测试划分：按时间顺序以前 80% 的 TR 作为训练集、后 20% 的 TR 作为测试集。该划分避免了 K 折交叉验证带来的

显著计算开销，使多模型多层多窗口的系统比较在当前硬件条件下可行。为保证不同模型之间的可比性，我们在所有模型上使用相同的划分方式与相同的正则化系数设置，并在多被试上重复该评估。对每个 ROI，我们计算测试集预测值与真实值的皮尔逊相关系数，得到该被试的 corr map；再对 corr map 的 ROI 维做均值得到该被试的平均相关。最终报告的均值与标准差来自对所有被试平均相关的汇总。

2.3.4 输出文件与可追溯性

为保证论文结论可核验，本文所有结果均对应到仓库中已生成的文件。单模态与多模态（Whisper/CLAP）结果以 results/GROUP/MODEL/SETTING/ 为目录结构，每个配置包含 log.txt 与若干 corr_layer*.npy，前者记录多被试均值与标准差，后者保存 ROI 级 corr map。融合结果保存到 results/fusion/TEXT_MODEL__AUDIO_MODEL/，每一组层与窗口配置保存为 corr_t*_a*_ctx*_tr*.npy 并在对应 log.txt 中记录统计量。统计图由 report/scripts/make_figures.py 从 results/summary.csv、results/roi.csv 与融合日志聚合生成到 report/figures/；皮层脑图由 src/run_plot_corr_maps.py 将 corr_*.npy 映射到皮层表面并输出到 report/figures/brainmaps/。这种“数值—可视化—文件路径”的对应关系保证了本文叙述可以被逐项复现与检查。

3

实验结果

3.1 实验结果（一）：文本模型

文本模型结果来自 `results/summary.csv` 中 `/text/` 条目及对应的 `corr_layer*.npy`。本次文本特征采用固定 200 token 上下文窗口，并将词级表示在 TR 内取平均得到 TR 级特征。该设置在方法上提供了稳定对照，但也意味着文本表征受限于窗口长度与 TR 内词数波动。由于本文不引入超出已完成实验的推断，结果部分仅对当前设置下的数值与空间分布作出可追溯陈述。

三种文本模型的最佳层均值相关分别为：RoBERTa-base 在 layer4 达到 0.0153 ± 0.0160 ，GPT-2 在 layer12 达到 0.0107 ± 0.0137 ，BERT-base 在 layer6 达到 0.0084 ± 0.0161 。图 3.1 以柱状图形式对比了三种模型的最佳层表现。就数量级而言，文本模型在本任务中的可预测性显著低于音频与多模态模型（后文将给出 0.08–0.09 量级的强基线），因此文本结果在本文中主要承担“纯语义/文本基线”的作用，用于在相同评估协议下刻画语义特征在当前数据与时间尺度下的可预测上限。

空间分布方面，图 3.2 将三种文本模型的最佳层相关图并置，便于观察在同一绘图视角与同一色标下的差异。为避免把跨模型差异误读为绘图设置差异，本图的每一张脑图都来自相同的 corr map 到皮层表面映射流程。进一步地，图 3.3、图 3.4 与图 3.5 分别展示三种模型在其最佳层的单独脑图，从而便于在后续讨论 ROI 偏好与空间模式时引用。尽管文本模型整体相关较低，图 3.6 的 ROI Top20 仍能为“哪些区域在当前框架下更容易被文本语义预测”提供定量入口，并可与音频/多模态的 ROI 分布形成对照。

3.2 实验结果（二）：音频模型

音频模型结果覆盖三种模型（wav2vec2、WavLM、HuBERT）以及四种 TR 窗口长度（1TR、2TR、3TR、6TR）。在音频侧，TR 窗口长度不仅改变了输入模型的时间上下文，也决定了 TR 级特征所汇聚的声学信息范围，因此它在自然故事听觉范式下具有直接的可解释意义。本文在统一的评估协议下比较不同窗口与不同层的表现，从而把“时间尺

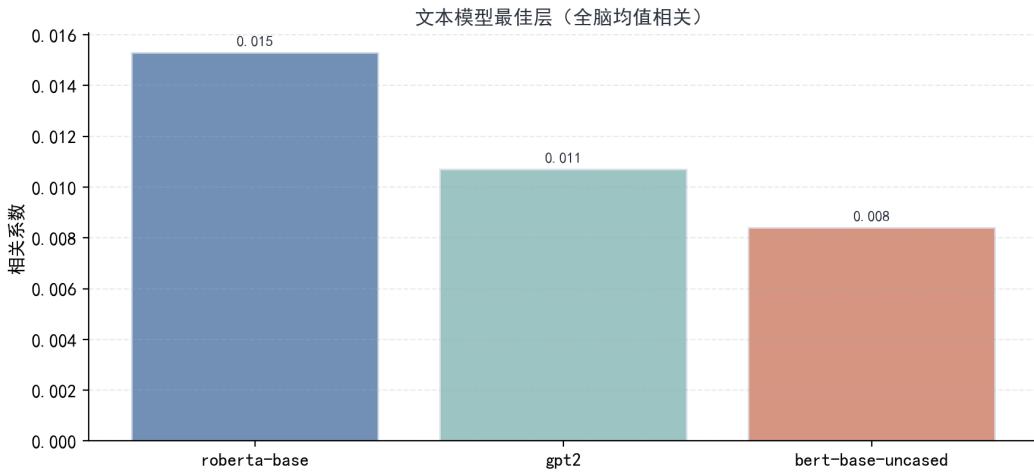


图 3.1 文本模型最佳层的全脑均值相关系数（从 `results/summary.csv` 聚合）。

度”作为与“层级位置”并列的系统变量。

从模型整体最佳表现看，WavLM-base-plus 的最优配置出现在 6TR、layer9，均值相关为 0.0916 ± 0.0405 ；wav2vec2-base-960h 的最优配置出现在 6TR、layer9，均值相关为 0.0895 ± 0.0320 ；HuBERT-base-ls960 的最优配置出现在 6TR、layer9，均值相关为 0.0823 ± 0.0346 。图 3.7 汇总了三种音频模型的最佳层表现，可以看到音频表征在当前数据与评估框架下构成全局强基线，其数量级显著高于文本模型。

TR 窗口效应在三种音频模型上表现为一致的单调提升。以“每个窗口内的最佳层”为代表，WavLM 在 1TR、2TR、3TR、6TR 的最佳均值相关分别为 0.0316 (layer9)、0.0450 (layer4)、0.0583 (layer9)、0.0916 (layer9)；wav2vec2 分别为 0.0274 (layer1)、0.0382 (layer1)、0.0569 (layer9)、0.0895 (layer9)；HuBERT 分别为 0.0326 (layer4)、0.0452 (layer4)、0.0556 (layer4)、0.0823 (layer9)。图 3.8 将这一趋势可视化为曲线，表明更长的声学上下文在当前编码框架下显著提高可预测性。这一现象与自监督语音表征“在更长上下文内形成更稳定结构表示”的观点相一致 [2, 6]，但本文在结果部分不将其解释为机制因果，仅将其作为在统一评估协议下的稳健经验结论。

空间分布方面，图 3.9 并置了三种音频模型在其最优配置下的脑图，从而在同一视角与色标下比较其皮层分布。图 3.10 展示全局最优音频配置（WavLM 6TR layer9）的单独脑图，图 3.11 与图 3.12 分别展示 wav2vec2 与 HuBERT 的最优配置脑图。ROI 层面，图 3.13 给出 WavLM 最优配置的 ROI Top20，为后续讨论“不同脑区对不同模态表征的偏好”提供对照基线。

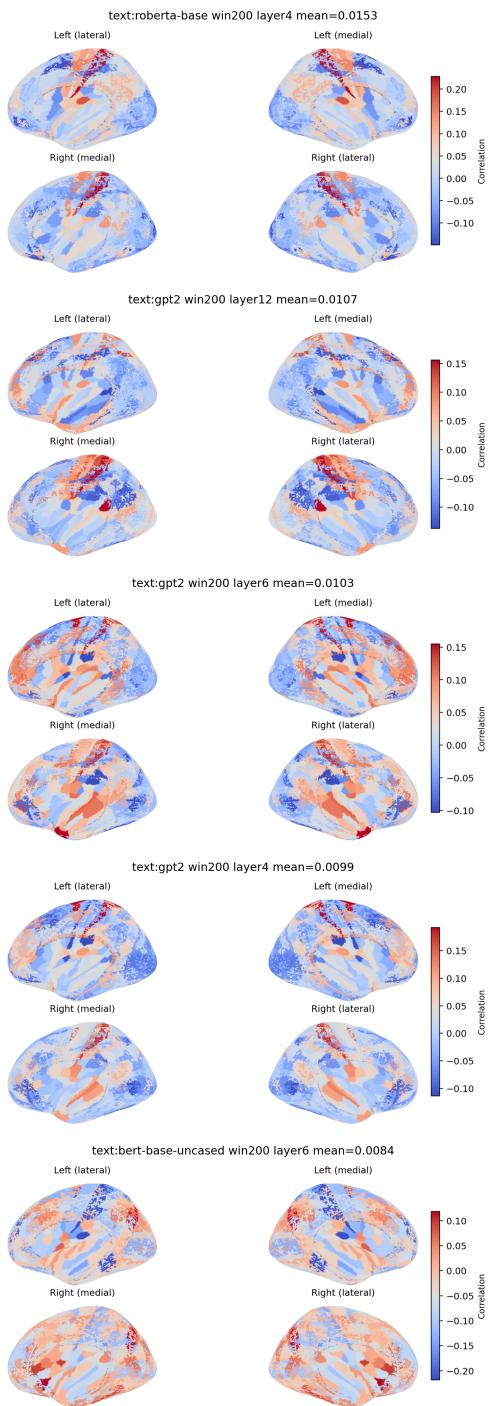


图 3.2 文本模型最佳层脑图对照：RoBERTa（win200, layer4）、GPT-2（win200, layer12）、BERT（win200, layer6）。

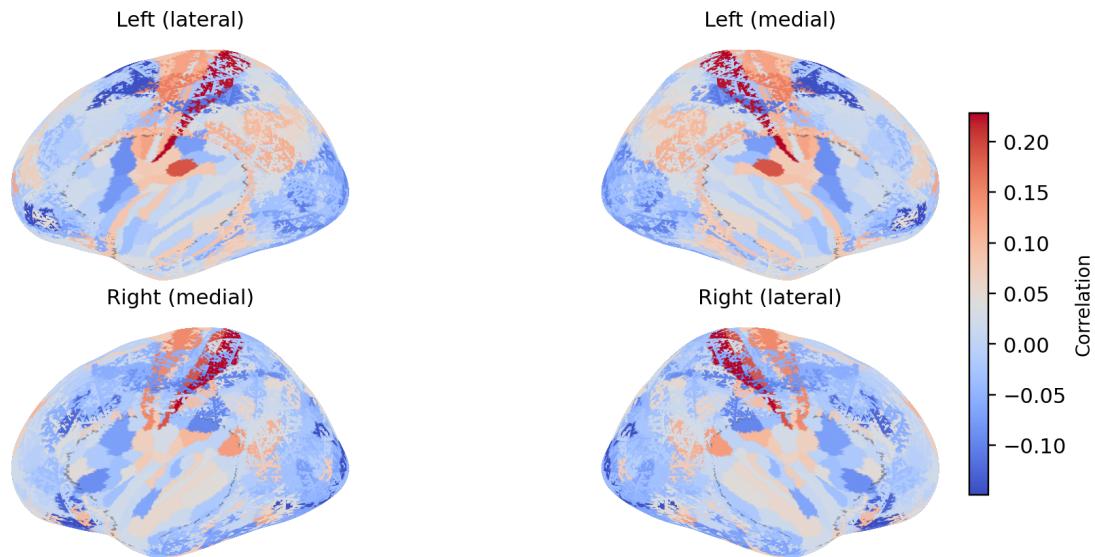


图 3.3 RoBERTa-base (win200, layer4) 相关图可视化。

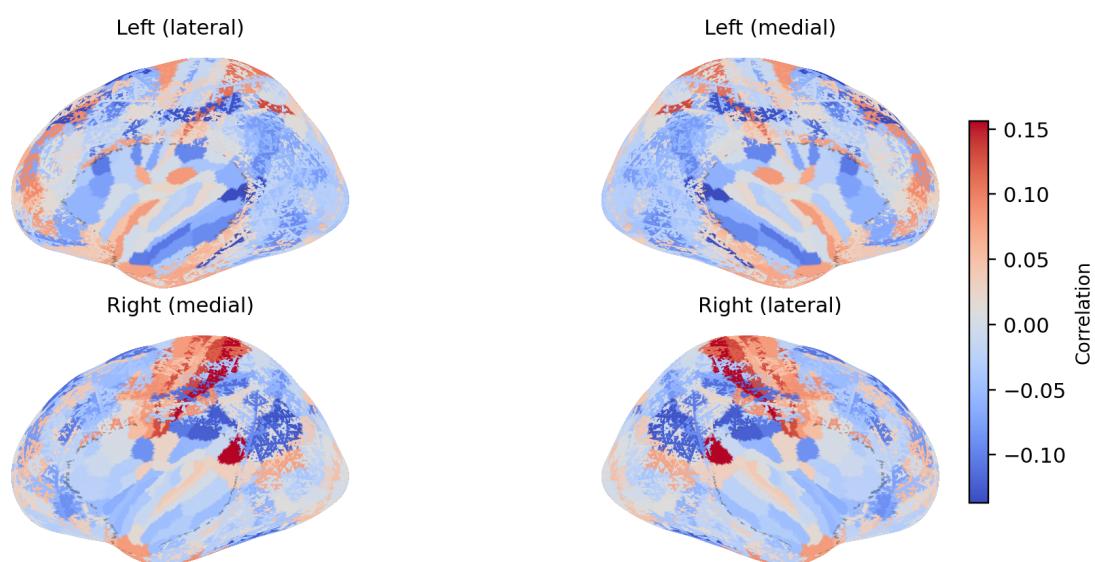


图 3.4 GPT-2 (win200, layer12) 相关图可视化。

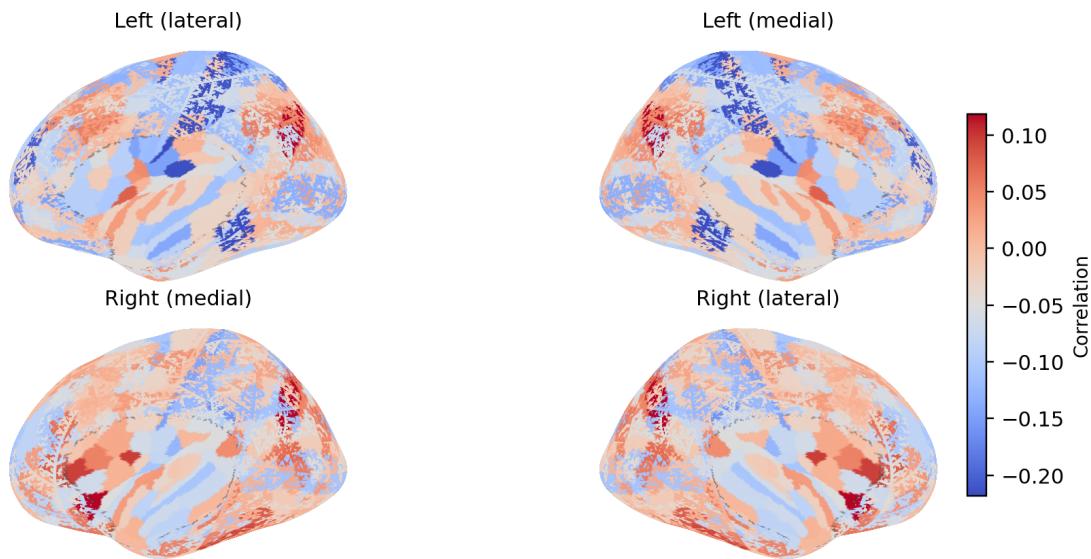


图 3.5 BERT-base (win200, layer6) 相关图可视化。

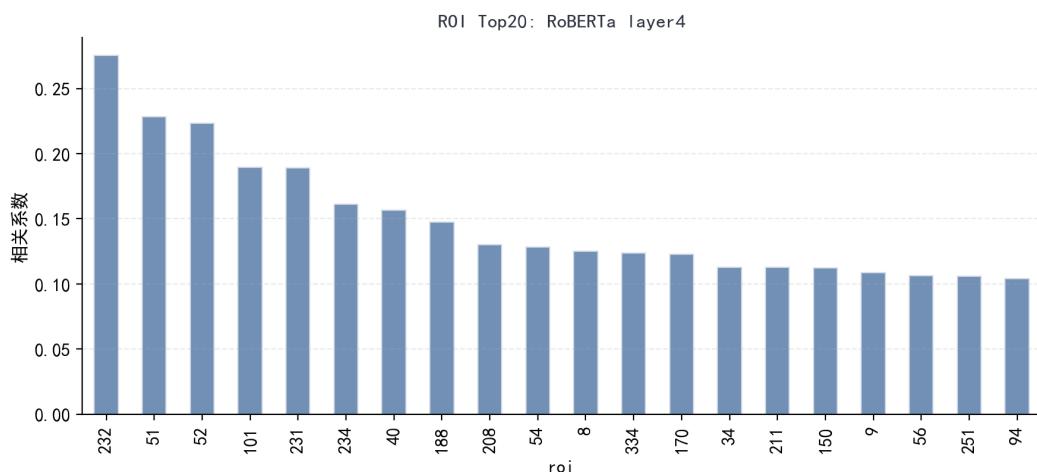


图 3.6 RoBERTa-base (win200, layer4) 对应相关图的 ROI Top20 (从 results/roi.csv 聚合)。

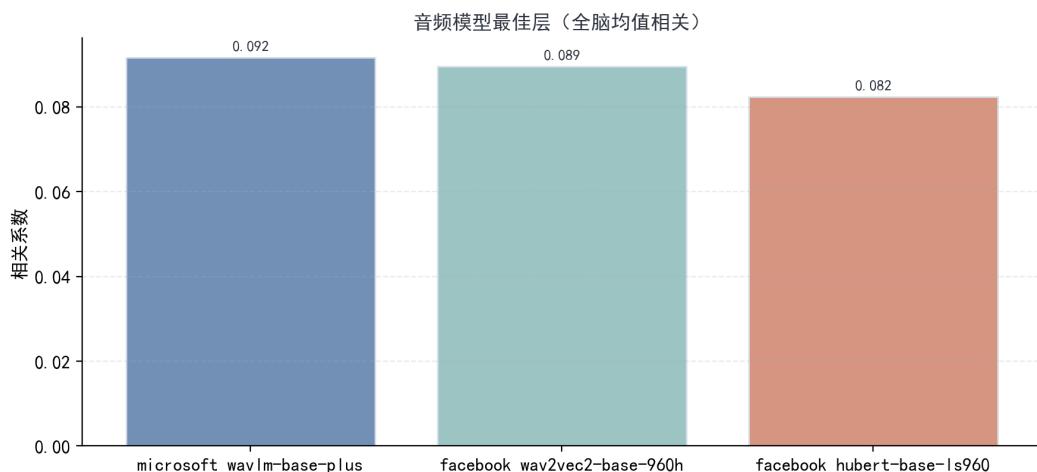


图 3.7 音频模型最佳层的全脑均值相关系数 (从 results/summary.csv 聚合)。

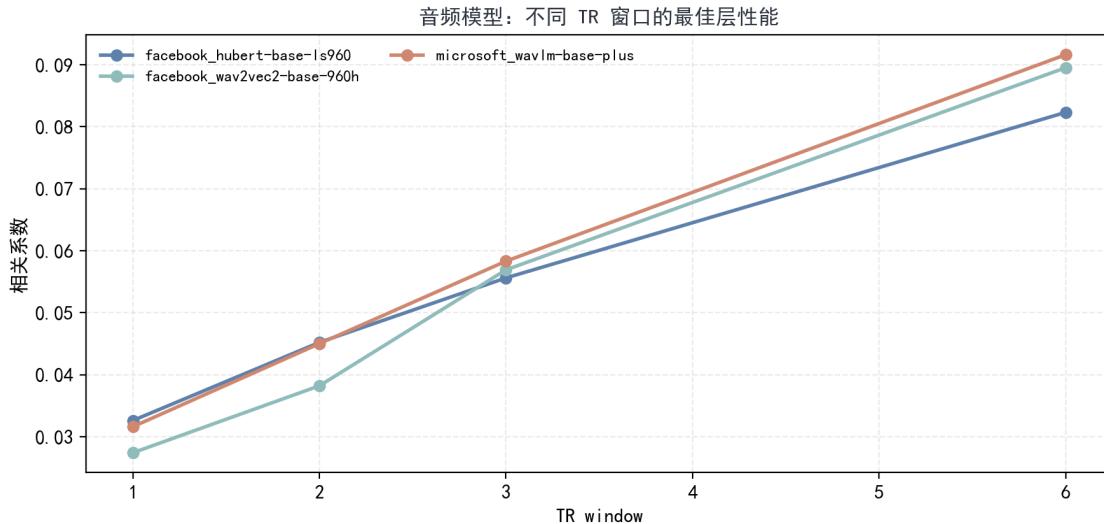


图 3.8 音频模型在不同 TR 窗口下的最佳层性能趋势（从 `results/summary.csv` 聚合）。

3.3 实验结果（三）：多模态模型

多模态模型结果来自 `results/summary.csv` 中 `/multimodal/` 条目及其对应的 `corr map`。与纯音频模型相比，多模态模型在训练目标或表示空间上显式引入了跨模态约束，因此它们在自然故事听觉任务中的表现能够回答一个具体问题：在统一的编码评估框架下，多模态表征是否在保持强音频基线的同时，进一步靠近语义系统的可预测模式。本文在当前已生成结果范围内对 Whisper 与 CLAP 的可用输出进行比较，并把其结果与音频基线对齐在同一量纲下呈现。

从整体最佳表现看，Whisper-base 的最优配置出现在 6TR、layer2，均值相关为 0.0889 ± 0.0321 ；Whisper-small 的最优配置出现在 6TR、layer9，均值相关为 0.0844 ± 0.0324 ；CLAP 的最优配置在当前结果中对应到 layer1，并在 6TR 时达到 0.0788 ± 0.0344 。图 3.14 汇总了多模态模型的最佳层表现，并可与图 3.7 的音频模型最佳层直接对照。就数值而言，Whisper-base 的最优均值与 wav2vec2/WavLM 的 6TR 最优结果处在同一量级，说明在自然故事听觉范式下，Whisper 的内部表征能够提供与强音频基线相近的可预测性，但并未显著超过最强音频模型。

TR 窗口效应在多模态模型上同样呈现一致的单调提升。以“每个窗口内的最佳层”为代表，Whisper-base 在 1TR、2TR、3TR、6TR 的最佳均值相关分别为 0.0270 (layer1)、0.0408 (layer4)、0.0579 (layer4)、0.0889 (layer2)；Whisper-small 分别为 0.0289 (layer9)、0.0417 (layer9)、0.0543 (layer12)、0.0844 (layer9)；CLAP 分别为 0.0231、0.0351、0.0509、0.0788 (均对应 layer1)。图 3.15 将这一趋势可视化，表明多模态模型同样依赖更长的声学

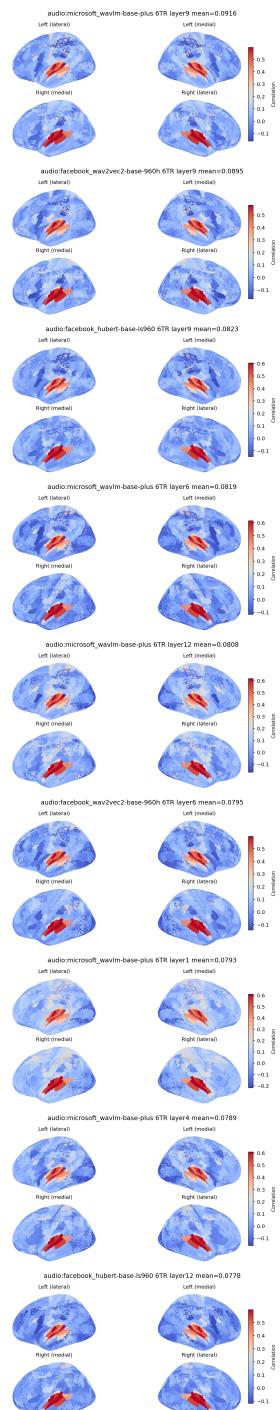


图 3.9 音频模型最优配置脑图对照：WavLM、wav2vec2、HuBERT。

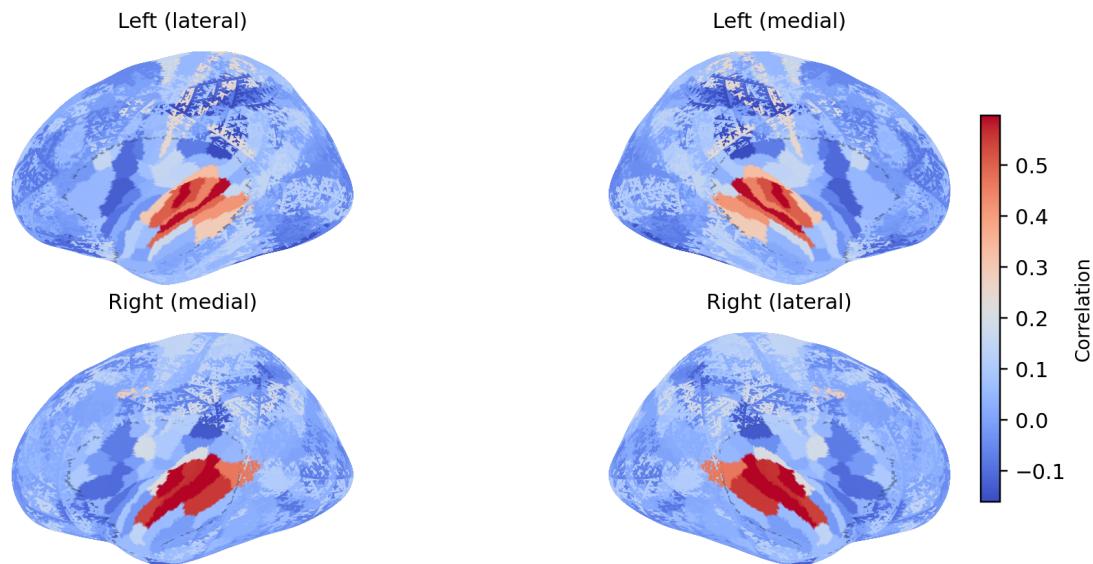


图 3.10 WavLM-base-plus (6TR, layer9) 相关图可视化。

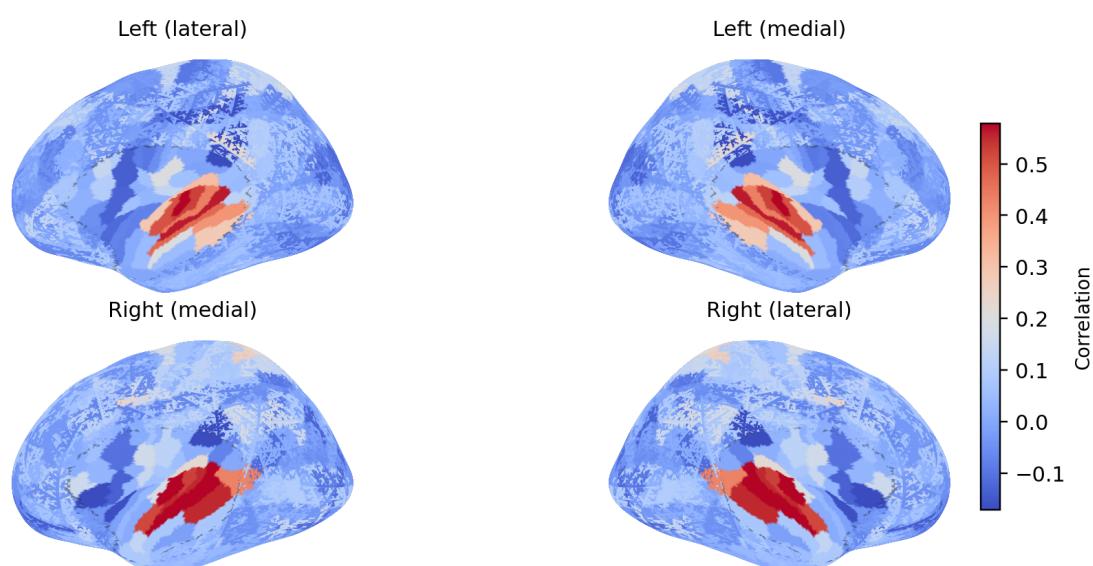


图 3.11 wav2vec2-base-960h (6TR, layer9) 相关图可视化。

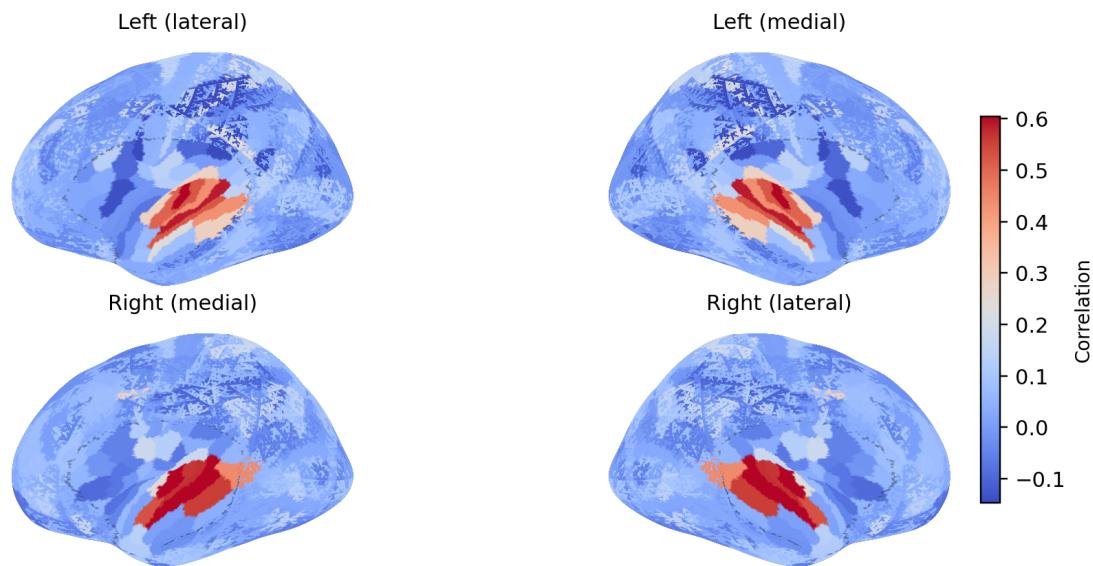


图 3.12 HuBERT-base-ls960 (6TR, layer9) 相关图可视化。

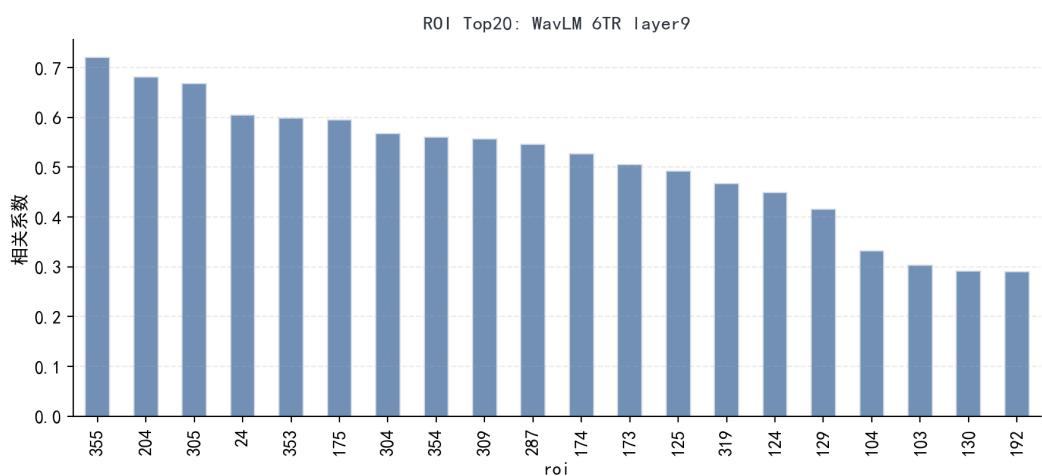


图 3.13 WavLM-base-plus (6TR, layer9) 对应相关图的 ROI Top20 (从 results/roi.csv 聚合)。

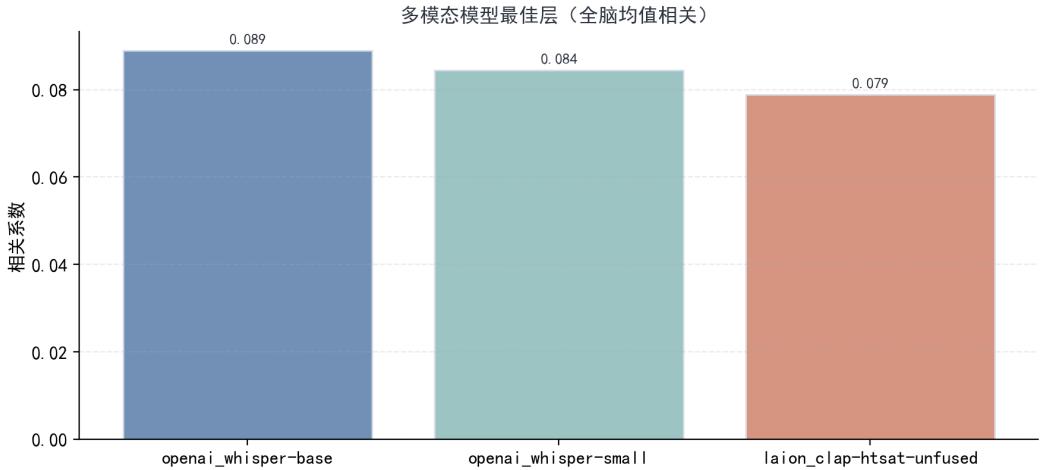


图 3.14 多模态模型最佳层的全脑均值相关系数（从 results/summary.csv 聚合）。

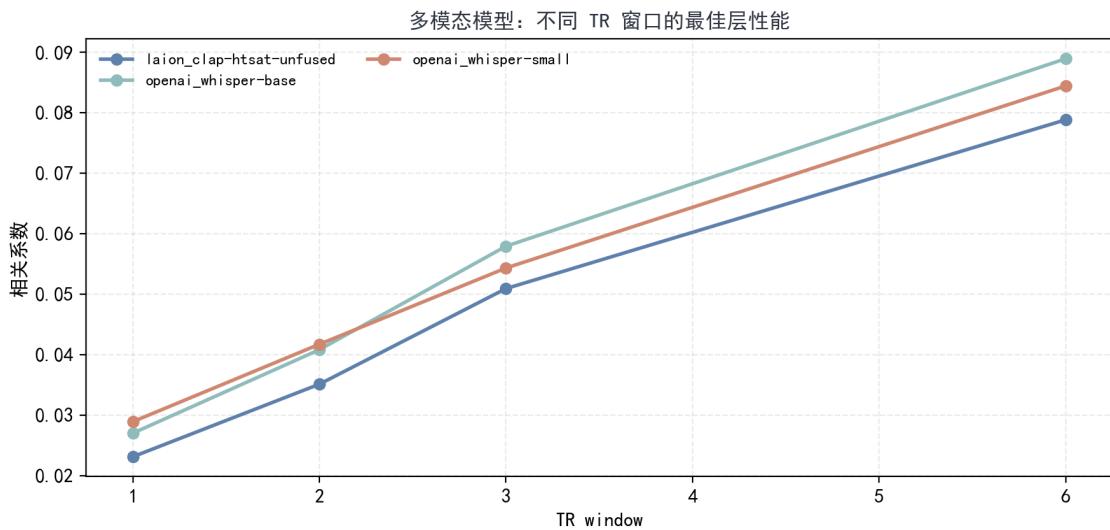


图 3.15 多模态模型在不同 TR 窗口下的最佳层性能趋势（从 results/summary.csv 聚合）。

上下文来提高对齐强度。结合自监督语音表征的层级对齐观点 [2, 6]，这一现象提示“时间尺度”在自然故事范式下对音频相关表征具有普遍影响，但其是否对应更高层语义整合仍需要在更丰富的语义控制实验中才能区分。

空间分布方面，图 3.16 并置了多模态模型的代表性脑图，用于与音频与文本结果在同一视角下比较。图 3.17 展示 Whisper-base 最优配置 (6TR, layer2) 的单独脑图，图 3.18 展示 Whisper-small 的最优配置 (6TR, layer9)。ROI 层面，图 3.19 给出 Whisper-base 最优配置的 ROI Top20，用于在讨论中分析其与音频基线在区域偏好上的相同与差异。需要说明的是，CLAP 的脑图在当前本地绘图目录中未形成可引用的图像文件，因此本文仅在数值层面对 CLAP 的可预测性进行报告，并将其空间分析留作后续补齐绘图输出后的扩展。

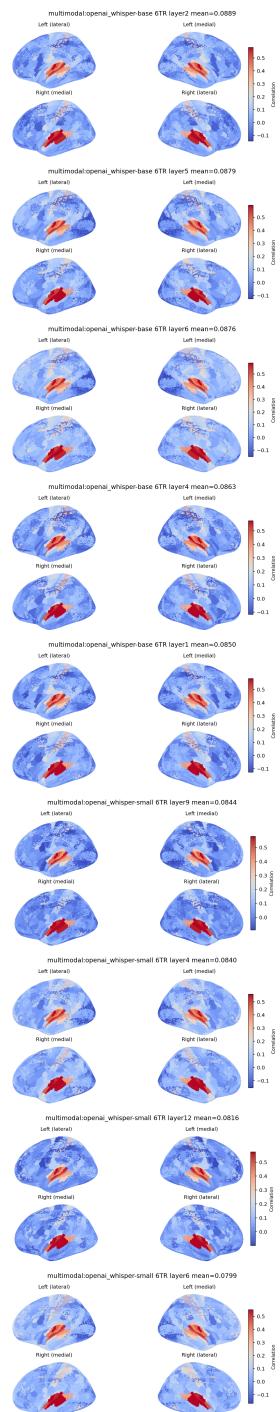


图 3.16 多模态模型代表性脑图对照（由本地已生成的脑图组合）。

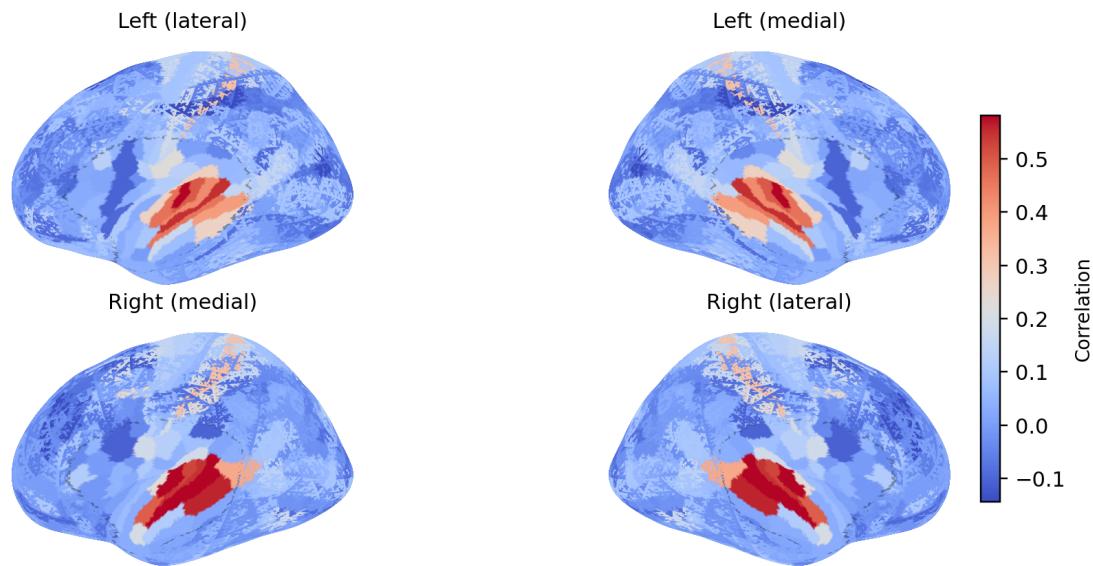


图 3.17 Whisper-base (6TR, layer2) 相关图可视化。

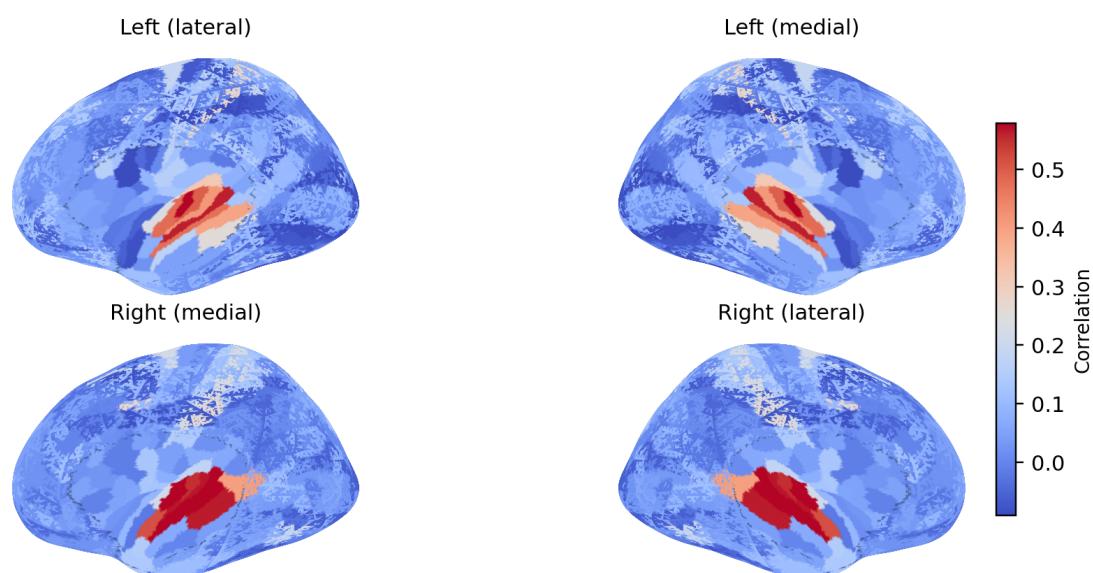


图 3.18 Whisper-small (6TR, layer9) 相关图可视化。

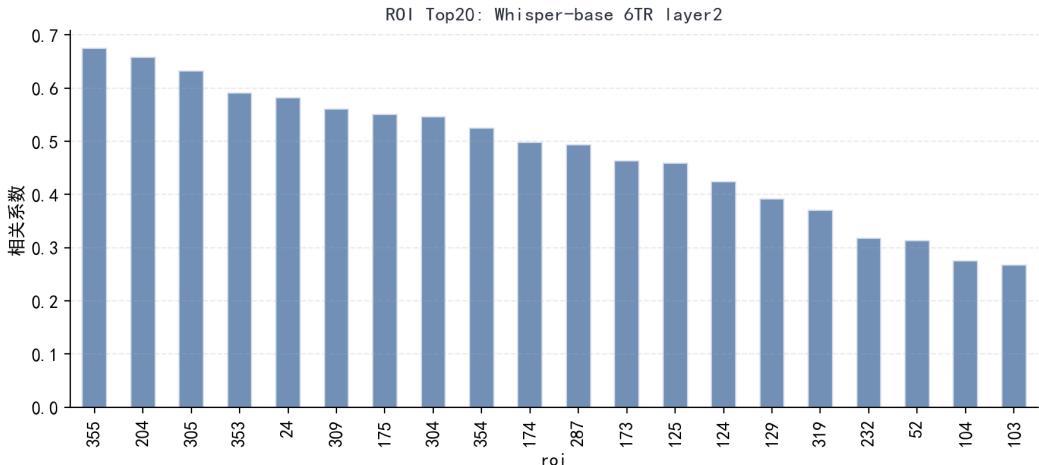


图 3.19 Whisper-base (6TR, layer2) 对应相关图的 ROI Top20 (从 results/roi.csv 聚合)。

3.4 实验结果 (四): 文本-音频特征融合

融合实验的目标是在不改变编码模型与评估协议的前提下，检验文本特征与音频特征是否在 TR 级时间轴上提供互补信息。与多模态模型“在预训练阶段引入跨模态对齐约束”不同，本文的融合采取更直接的特征级拼接：对同一 TR 的文本特征与音频特征分别标准化后在特征维上拼接，再在拼接后的联合空间执行 PCA，随后进入同一 FIR+ 岭回归评估流程。该实现对应 results/fusion/ 目录，其中文本与音频模型对以子目录区分，具体配置以 corr_t*_a*_ctx*_tr*.npy 命名并在 log.txt 中记录统计量。

当前已生成的融合结果覆盖固定文本上下文窗口 $\text{ctx} = 200$ 、三种音频 TR 窗口 (1TR、2TR、3TR)、三种文本模型与三种音频模型，并在多层组合上形成可解析的配置集合。以融合日志解析得到的 540 条记录为基础，融合的全局最优配置出现在 $\text{tr} = 3$: RoBERTa-base 的 $\text{text_layer} = 1$ 与 WavLM-base-plus 的 $\text{audio_layer} = 9$ 组合达到 0.0535 ± 0.0266 。分窗口看， $\text{tr} = 1$ 的最优配置为 RoBERTa-base (layer1) + HuBERT (layer4)，均值相关为 0.0330 ± 0.0179 ; $\text{tr} = 2$ 的最优配置为 RoBERTa-base (layer1) + WavLM (layer9)，均值相关为 0.0431 ± 0.0212 ; $\text{tr} = 3$ 的最优配置即上述全局最优。图 3.20 与图 3.21 分别展示融合 Top 配置的整体对比与窗口趋势，清晰呈现“窗口增大带来系统性提升”的规律。

融合结果的关键问题并不仅是“是否提高一个全局均值”，而是“文本层与音频层是否呈现交互结构”。若融合收益仅由某一侧单模态强信号驱动，则不同层组合应当在二维网格上近似单调；相反，若存在互补与匹配，则可能出现局部最优区域。图 3.22 在固定 $\text{ctx} = 200$ 、 $\text{tr} = 3$ 与全局最优模型对 (RoBERTa + WavLM) 条件下给出二维性能热图，

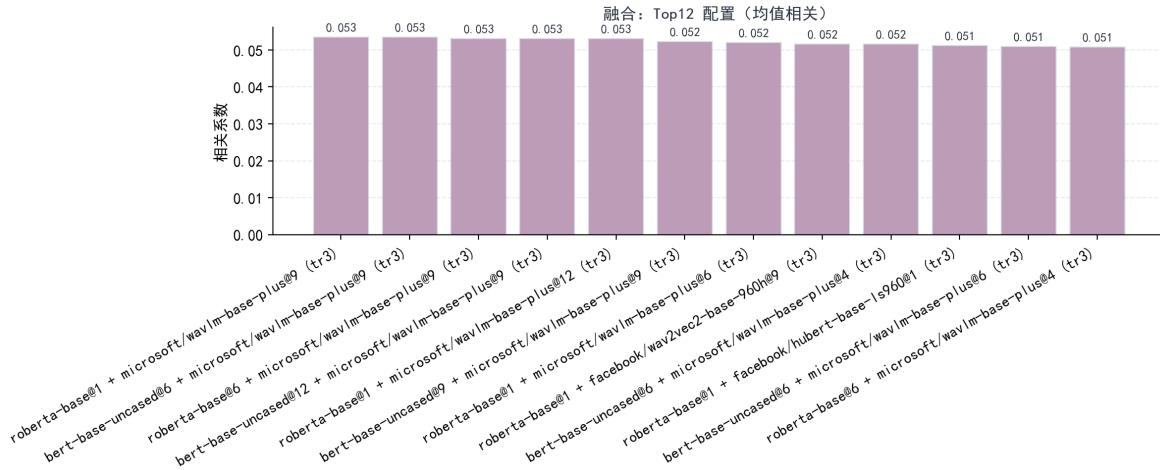


图 3.20 融合：Top 配置的全脑均值相关对比（从 `results/fusion/**/log.txt` 解析聚合）。

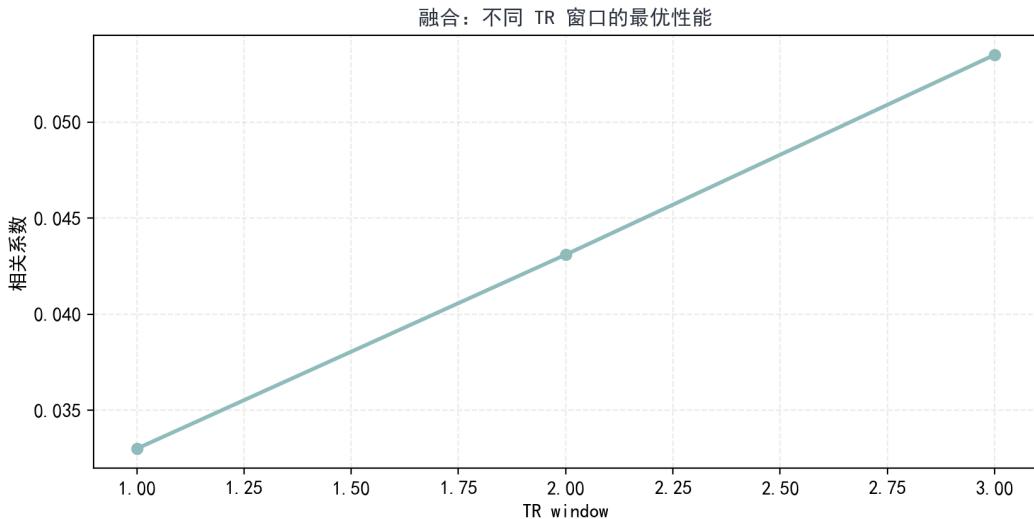


图 3.21 融合：不同 TR 窗口下的全局最优性能趋势（从融合日志解析聚合）。

可以看到性能分布并非简单随层号单调变化，而是在若干组合附近形成高值区域。这一现象与“层级位置决定表征抽象度与可用信息类型”的观点一致 [9, 1, 5]，并为后续更系统的跨层融合策略提供了直接的经验约束。

空间分布方面，图 3.23 展示融合 Top 配置的脑图对照，图 3.24 展示全局最优融合配置的单独脑图，从而把融合结果纳入与单模态一致的“统计图—ROI 图—脑图”证据链。需要强调的是，融合最优均值相关（约 0.053）仍明显低于音频与多模态在 6TR 条件下的 0.08–0.09 量级强基线，因此在当前结果覆盖范围内，特征拼接融合并未带来超越强音频模型的整体优势。该结论并不意味着跨模态互补不存在，而更可能反映当前融合只覆盖到 3TR 的时间尺度、以及联合 PCA 在有限样本下对跨模态方差结构的选择性等因素。对这些因素的分析将放在讨论部分展开。

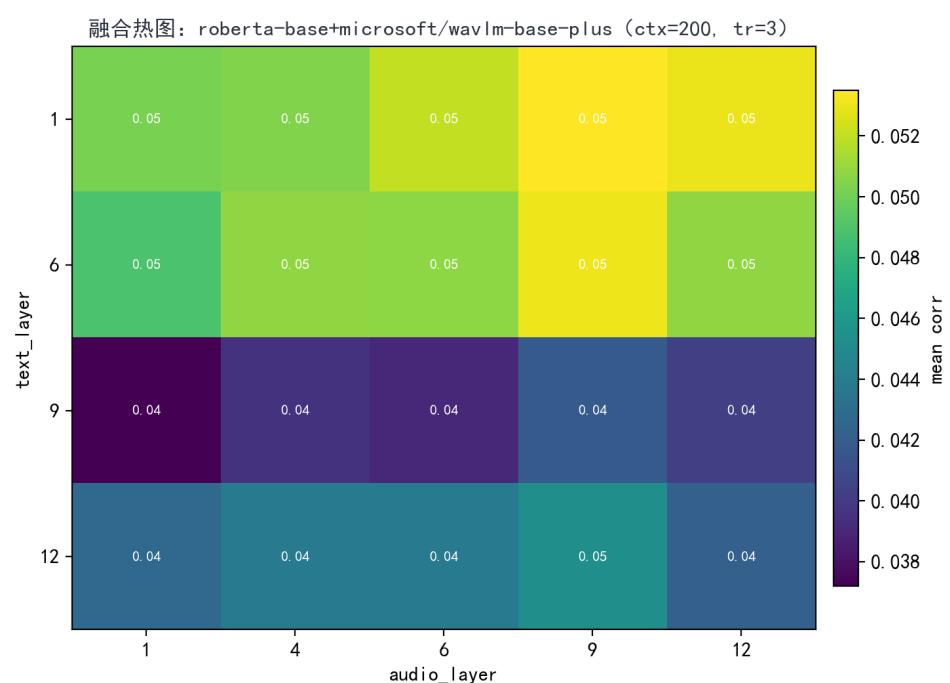


图 3.22 融合热图: RoBERTa + WavLM 在 $\text{ctx} = 200$ 、 $\text{tr} = 3$ 条件下的层交互结构 (从融合日志解析生成)。

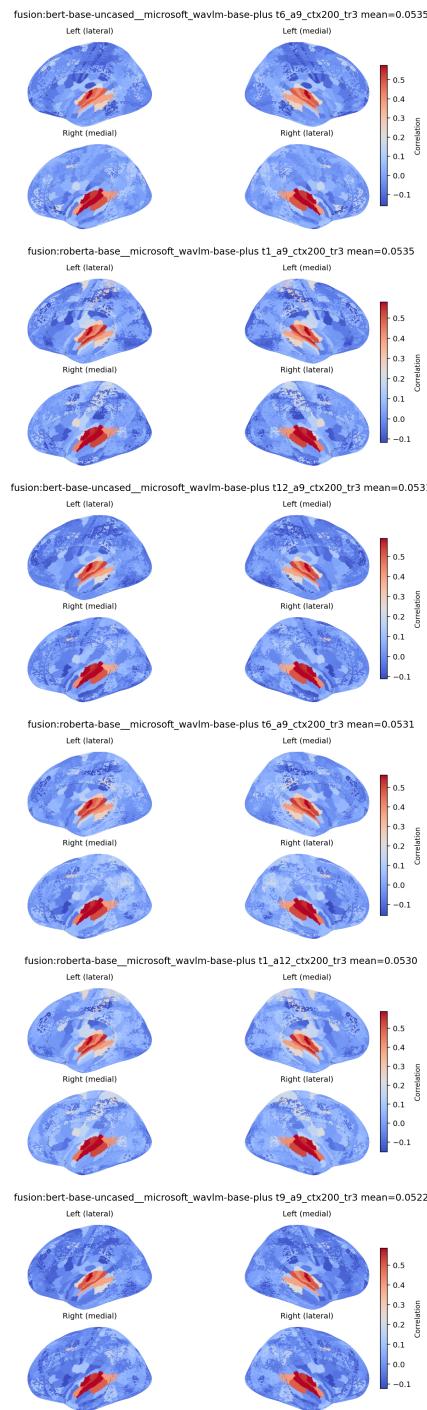


图 3.23 融合脑图对照：融合 Top 配置的 corr map 脑图组合。

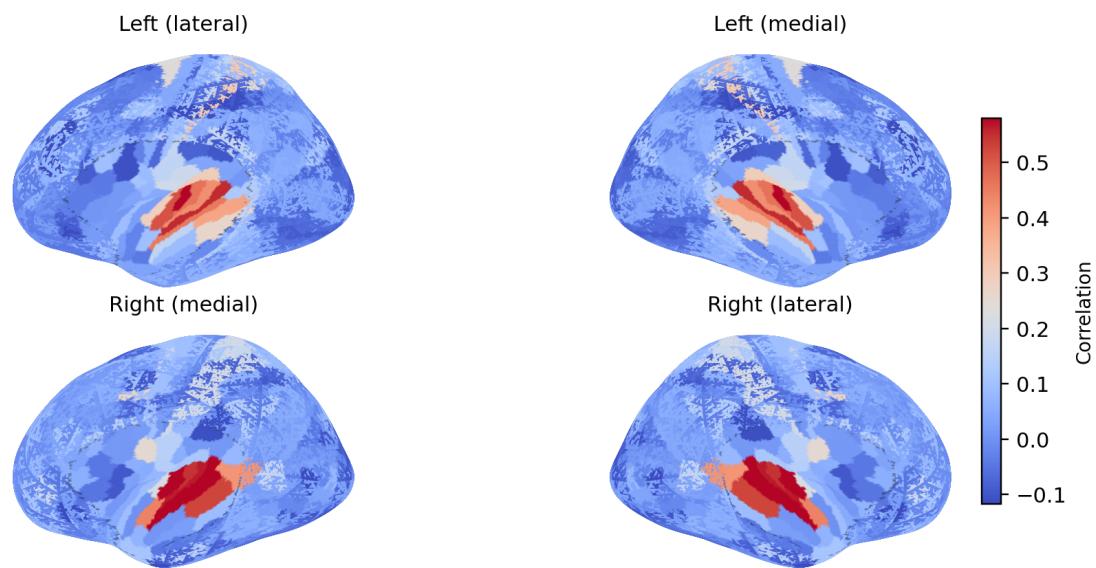


图 3.24 全局最优融合配置脑图: RoBERTa-base (layer1) + WavLM-base-plus (layer9), $\text{ctx} = 200$, $\text{tr} = 3$ 。

4

讨论

4.1 讨论：结果的意义、不足与展望

本文在统一的编码建模与评估协议下，比较了文本、音频、多模态与特征级融合四类表征对自然故事听觉 fMRI 的可预测性，并以 ROI 统计与皮层脑图给出可视化证据链。讨论部分围绕三个问题展开：第一，这些结果对“语义系统的可预测性与分布式组织”意味着什么；第二，结果中最显著的时间尺度效应应如何理解；第三，在当前实现边界内，哪些不足限制了结论外推，后续应如何扩展以更接近研究动机中提出的问题。

从对齐强弱的全局排序看，音频与多模态模型在 6TR 条件下达到约 0.08–0.09 的均值相关，而文本模型的最佳配置仅约 0.015。这一差异提示，在自然故事听觉范式下，编码模型在当前设置中更容易利用声学与语音相关信息来预测 BOLD 变化，而纯文本语义表征在相同时间对齐方式下贡献较弱。该结论并不与语义地图研究矛盾。Huth 等的语义地图建立在“明确构造的语义特征空间”与“排除或控制声学协变量”的设计之上，并且其特征与任务设置更直接对准语义维度 [4]；Zhang 等对语义关系的映射同样依赖对语义类别、具体性与关系向量的针对性分析 [12]。相比之下，本文的文本特征来自通用预训练语言模型的隐藏状态，其语义信息与句法信息、词形信息及上下文统计规律混合在同一高维空间中，且未在特征层面显式控制声学因素。这意味着“文本模型相关较低”更可能反映在当前对齐与回归设置下，语义信息被多种因素稀释或被声学驱动信号掩盖，而不是意味着语义系统不存在或不可预测。

TR 窗口长度对音频、多模态与融合结果的单调提升，是本文最稳定也最具解释张力的现象。对音频模型而言，窗口从 1TR 增至 6TR 时均值相关从约 0.03 提升到约 0.09，且该趋势在 wav2vec2、WavLM 与 HuBERT 上一致。多模态模型也呈现同样趋势，Whisperbase 从 1TR 的约 0.027 提升到 6TR 的约 0.089，CLAP 也从约 0.023 提升到约 0.079。这说明在自然故事听觉范式下，短窗表征不足以提供对 BOLD 变化稳定可用的线性预测信号，而更长时间范围内的汇聚显著提高了可预测性。自监督语音模型的训练目标本就鼓励模型在更长上下文内整合信息，例如 wav2vec 2.0 的掩码预测需要依赖上下文来恢复被遮蔽片段 [2]，而与脑对齐研究指出其层级结构可与皮层语音处理层级对应 [6]。因此，窗口效

应既可能反映模型表征对长程声学统计规律的利用，也可能反映 BOLD 动力学下“可被线性模型捕捉的有效信号”需要更长时间汇聚才能显现。由于本文在 FIR 展开中使用固定窗口与偏移，这两种因素在结果上并未被分离，后续需要更系统地在特征窗口与 FIR 参数上做正交控制，才能区分“输入汇聚”与“动力学建模”各自的贡献。

ROI 统计为“空间偏好”提供了更细粒度的证据。图 4.1、图 4.2、图 4.3 分别展示 RoBERTa 文本基线、WavLM 强音频基线与 Whisper-base 强多模态配置的 ROI Top20。尽管本文的 ROI 分区来自固定脑区划分而非体素级地图，但仍可用于观察不同表征在不同功能区域的相对优势。文本基线的 Top ROI 与音频/多模态基线的 Top ROI 之间既有重叠也存在差异，这与语义系统“跨网络分布式组织”的观点一致 [4, 12]：同一语义处理过程并不局限在传统语言区，而是涉及默认模式网络、顶叶与颞叶的协同活动。进一步地，Transformer 层级与功能分化分析提示模型内部不同层可能对应不同类型的信息整合，因而其可预测性在不同脑区的分布也可能呈现梯度 [5]。本文当前的 ROI 分析只展示了少数代表配置的 Top20 统计，后续可在保持评估协议不变的前提下，对更多层与更多模型绘制 ROI 分布并检验其稳定性，从而把“空间偏好”从单点观察扩展到系统规律。

融合实验提供了一个与多模态模型不同的对照：它不依赖跨模态预训练约束，而是在特征层直接拼接并在联合空间做 PCA。结果显示融合最优配置在 3TR 达到约 0.053，仍显著低于 6TR 条件下的音频与多模态强基线。该现象不应被简化为“文本无用”，因为融合覆盖的时间尺度仅到 3TR，且联合 PCA 可能把文本中较弱但互补的方向压缩掉。更重要的是，融合热图显示层组合存在明显交互而非单调趋势，这与集成建模对“层选择关键性”的强调一致 [9]，也与 Antonello 与 Huth 提出的“对齐来源可能来自更一般的结构归纳”相呼应 [1]：若两种表征空间在某些抽象层级上更匹配，则简单拼接也可能在局部产生收益。要把这一线索推进为更强结论，需要在更长 TR 窗口下补齐融合覆盖，并把融合策略从简单拼接扩展到更明确的跨模态对齐方式，例如在不改变编码模型的前提下对两种特征做对齐子空间学习或分块正则化。

本文也存在明确的不足与边界。其一，本文采用线性岭回归并以单次训练/测试划分在多被试上汇总，避免了交叉验证开销，但也意味着模型选择与不确定性估计较为保守。其二，文本侧上下文窗口固定为 200 token，模型端池化与 TR 内池化的组合在当前结果中并未做系统调参，因而无法回答“更长上下文是否提升语义对齐”这类问题；组合语义研究表明 supra-word 表征依赖精细的控制与比较 [11]，而本文目前的文本设置更接近基线而非优化。其三，多模态模型部分虽然包含 Whisper 与 CLAP 的数值结果，但空间可视化与更完整的层级对比仍需补齐；此外，本文未包含连续语义重构那样的解码结果 [10]，因此讨论严格限定在编码框架内。其四，ROI 级分析虽然提高了稳定性与可解释性，

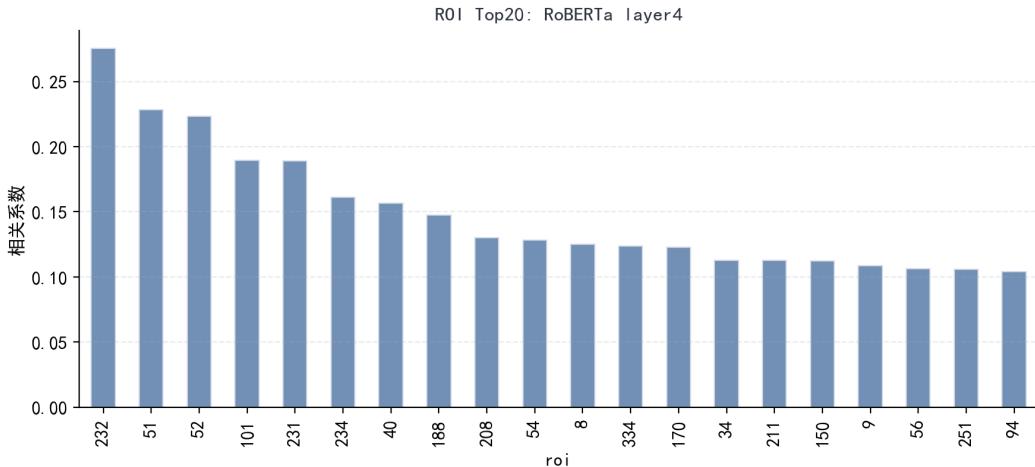


图 4.1 ROI Top20: RoBERTa-base (win200, layer4)。

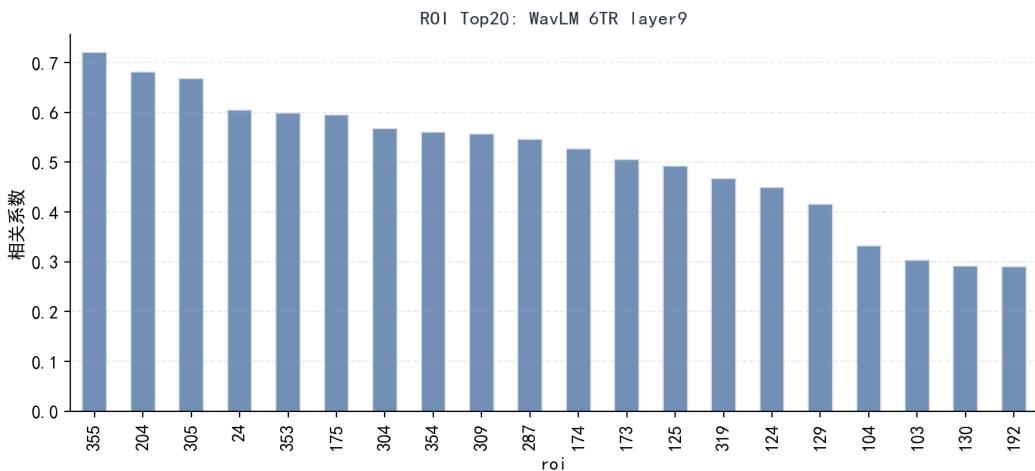


图 4.2 ROI Top20: WavLM-base-plus (6TR, layer9)。

但会平滑掉体素级差异，后续若要更接近语义地图工作中的精细分区，需要在体素级或表面顶点级上复现相同评估逻辑 [4]。

在上述边界内，本文给出三条可直接落地的扩展方向。第一，在保持评估协议不变的前提下，对文本上下文窗口、token 池化策略与 TR 内聚合方式做正交比较，从而把文本基线从“固定设置”推进到“系统调参后的强基线”，并检验其是否接近语义地图工作揭示的语义系统组织。第二，对音频与多模态模型进一步系统化“窗口长度—层级位置”的二维比较，并结合 ROI 分布检验层级梯度是否与皮层层级加工相对应 [6, 5]。第三，在融合框架下补齐更长 TR 窗口与更完整的层覆盖，结合热图交互结构探索更合理的跨模态融合策略，并在不引入额外假设的情况下检验“互补信息是否存在且位于哪些层级位置”。这些扩展能够在不改变数据与总体评估框架的前提下，逐步把本文从“可复现的系统比较”推进到“对机制更有约束力的证据链”。

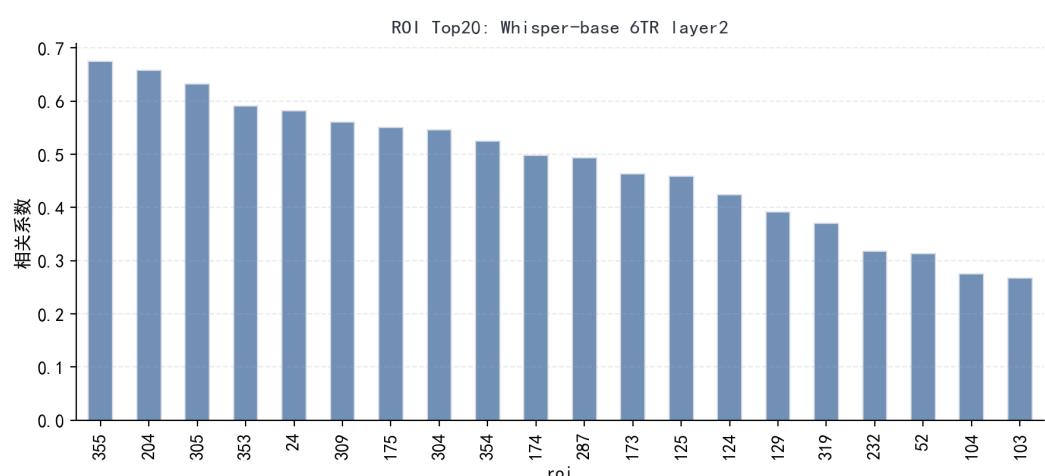


图 4.3 ROI Top20: Whisper-base (6TR, layer2)。

5

结论

5.1 结论

本文以自然故事听觉范式下的多被试 fMRI 数据为对象，在统一的对齐、降维与时延展开流程下构建线性编码模型，系统比较了文本模型、音频模型、多模态模型及文本-音频特征拼接融合对脑信号的可预测性。结果在当前已生成文件范围内呈现出清晰的经验结构：文本模型在固定 200 token 上下文窗口与 TR 内平均聚合的设置下提供了较弱但可追溯的语义基线；音频模型与多模态模型在更长 TR 窗口下显著提高可预测性，并在 6TR 条件下达到约 0.08–0.09 的均值相关；融合在覆盖到 3TR 的范围内呈现稳定的窗口提升与层级交互结构，但其最优均值仍低于强音频/多模态基线。ROI 统计与皮层脑图可视化进一步表明，不同表征在空间分布上既有共性也存在差异，为后续围绕语义系统分布式组织与模型层级功能分化的分析提供了可视化基础 [4, 12, 5]。

在方法层面，本文的贡献在于给出一条可复现的证据链：每个配置的数值统计、corr map 与图像输出都可以在仓库中逐项对应并核验，从而使跨模型比较建立在统一评估协议之上。与此同时，本文也明确了当前边界：线性编码与单次训练/测试划分限制了对不确定性的估计；文本侧上下文窗口与池化策略尚未系统调参；融合覆盖的时间尺度与层覆盖仍有限；多模态模型的空间分析仍需在绘图输出更完整后补齐。

在后续工作中，最直接的扩展路径是围绕“时间尺度”与“层级位置”做更充分的正交比较，并在不改变数据与总体评估逻辑的前提下提升文本与融合基线的强度。结合相关工作中的语义地图、组合语义与自监督语音层级对齐研究，这些扩展有望把本文当前的系统比较推进到对“哪些信息类型在何种时间尺度与层级位置上更接近可测脑信号”更有约束力的结论 [9, 1, 6, 11]。

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 Tetraeault, 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.