

认知神经科学

语义编码与多模态对齐：实验报告

基于故事听觉 fMRI 的预训练模型特征比较

Semantic encoding and multimodal alignment: an 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

本实验报告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。最终结论以当前“已经完成”的实验为依据：音频表征在长窗口下显著优于文本表征；Whisper 的多模态表征与强音频基线接近；文本 + 音频融合目前仅完成部分组合与窗口，尚未显示超过强音频基线的优势。

目录

1	背景、目标与本报告范围	3
2	数据、对齐表与 TR 级刺激构建	4
3	预训练模型、层选择策略与特征提取实现	5
3.0.1	模型集合与本次报告覆盖范围	5
3.0.2	层选择：按相对深度等比例取样	5
3.0.3	文本特征：上下文窗口与双层池化	5
3.0.4	音频特征：TR 窗口切分与帧级池化	6
3.0.5	多模态特征：模型内部融合与 TR 对齐	6
4	编码模型、评价指标与输出结构	6
4.0.1	从 TR 特征到 BOLD：线性编码模型的形式化	6
4.0.2	BOLD 延迟建模：PCA 降维与 FIR 延迟展开	7
4.0.3	数据划分与多被试汇总	7
4.0.4	评价指标与 corr map 的含义	7
4.0.5	输出文件结构与可追溯性	8
5	文本模型结果：层次对齐与区域分布	8
6	音频模型结果：TR 窗口效应与最佳层比较	9
7	多模态模型结果：Whisper	12
8	文本 + 音频融合结果：覆盖范围、最优配置与层交互结构	19
9	ROI 分析与综合讨论	21
10	结论、局限与后续可扩展方向	24
A	附录：结果文件位置与复现说明	25

1

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。

2

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 $\lceil t/\text{TR} \rceil$ 。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。

3

预训练模型、层选择策略与特征提取实现

3.0.1 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。

3.0.2 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。

3.0.3 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。

3.0.4 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。

3.0.5 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。

4

编码模型、评价指标与输出结构

4.0.1 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。

4.0.2 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。

4.0.3 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。

4.0.4 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。

4.0.5 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。

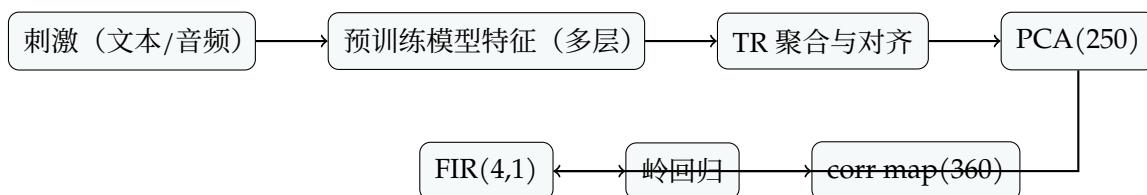


图 1 线性编码建模流水线概览。为避免版面溢出，示意图采用两行布局，但顺序与实现一致。

5

文本模型结果：层次对齐与区域分布

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。

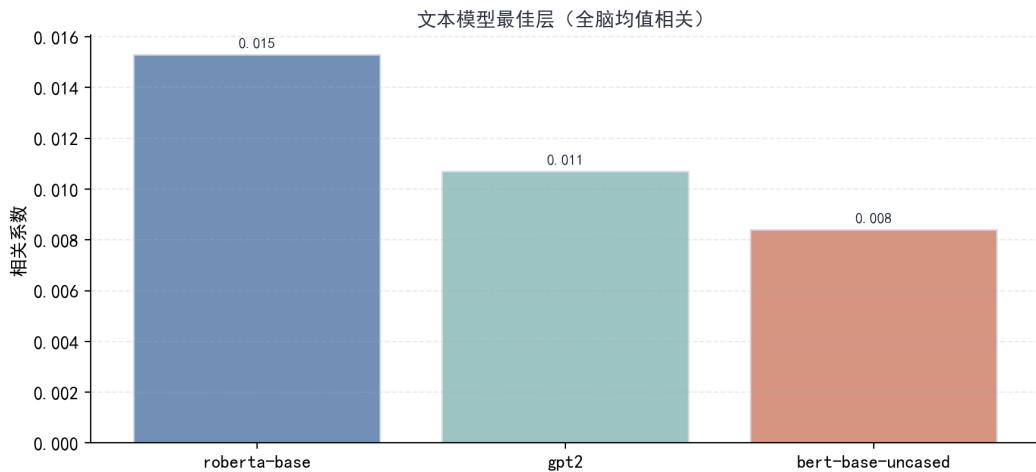


图 2 文本模型最佳层的全脑均值相关系数（由 report/scripts/make_figures.py 从 results/summary.csv 生成）。

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。

6 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。

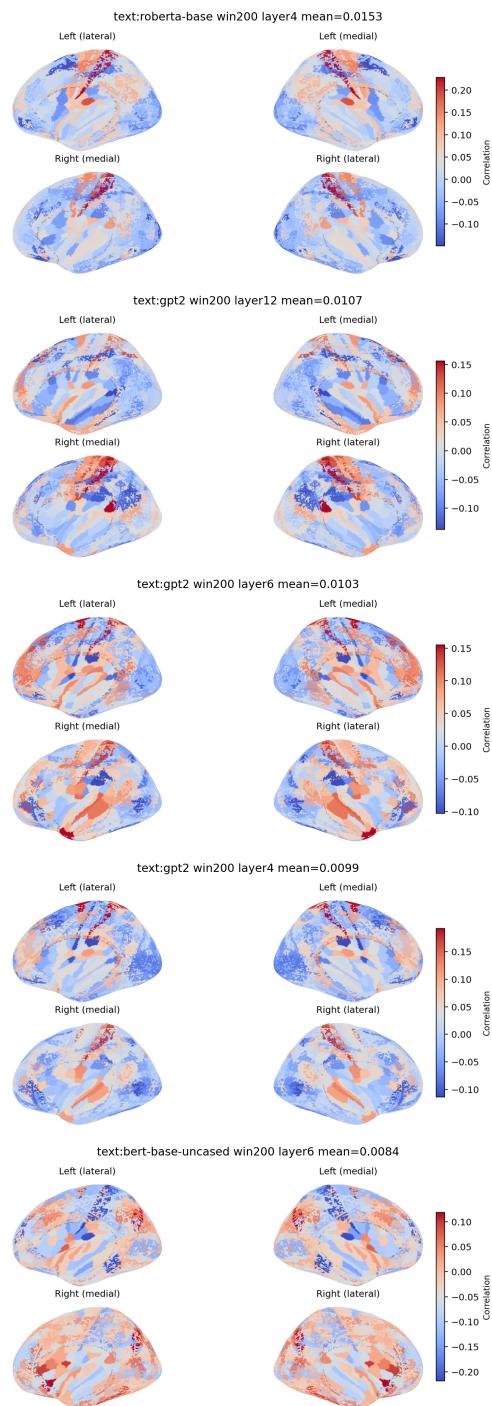


图3 文本模型类别脑图对照：RoBERTa (win200, layer4)、GPT2 (win200, layer12)、BERT (win200, layer6) (由 src/run_plot_corr_maps.py 从对应 corr_layer*.npy 绘制并组合)。

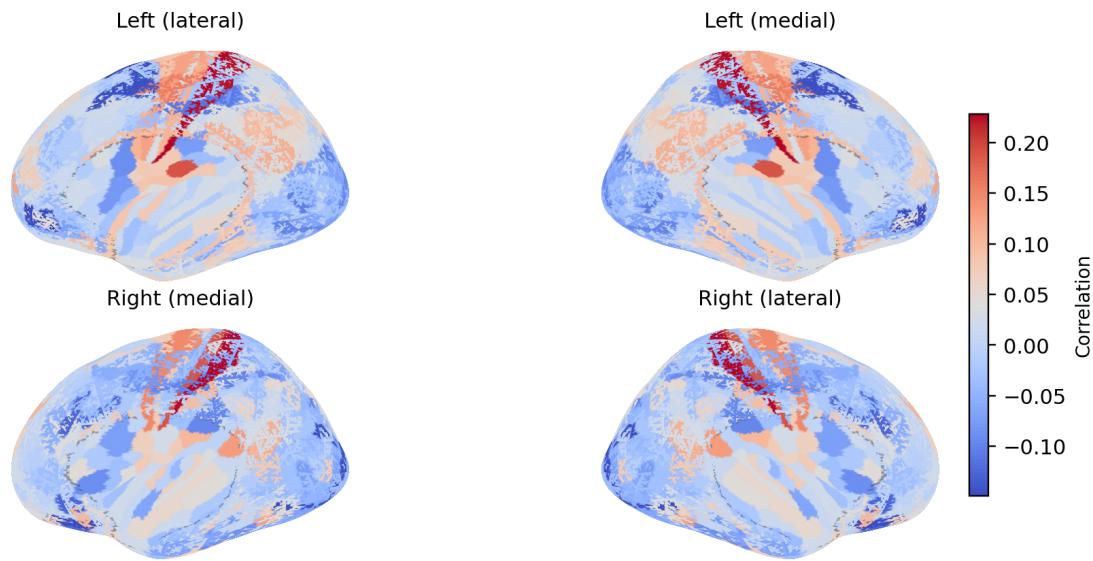


图 4 RoBERTa-base (win200, layer4) 相关图可视化（由 `src/run_plot_corr_maps.py` 从 `corr_layer4.npy` 绘制）。

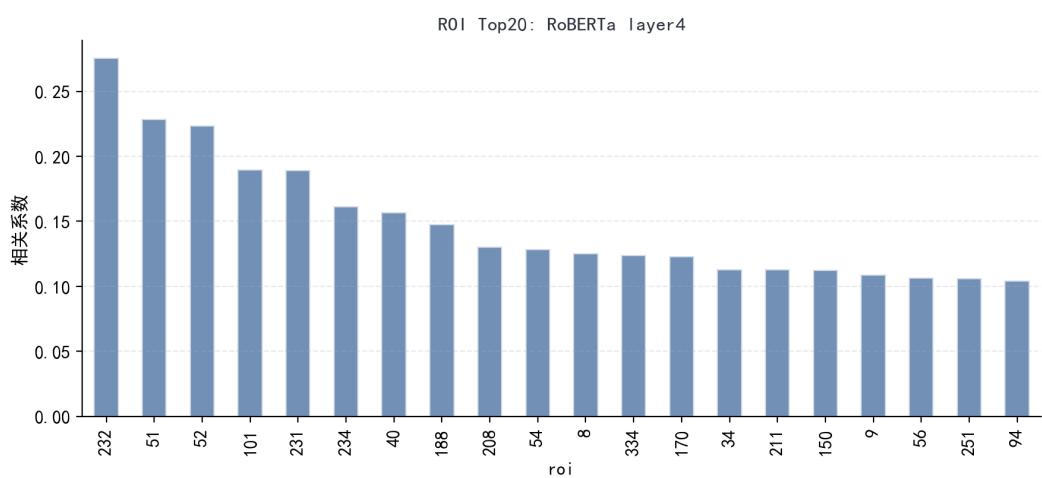


图 5 RoBERTa-base (win200, layer4) 对应 corr map 的 ROI Top20 (由 `report/scripts/make_figures.py` 从 `results/roi.csv` 生成)。

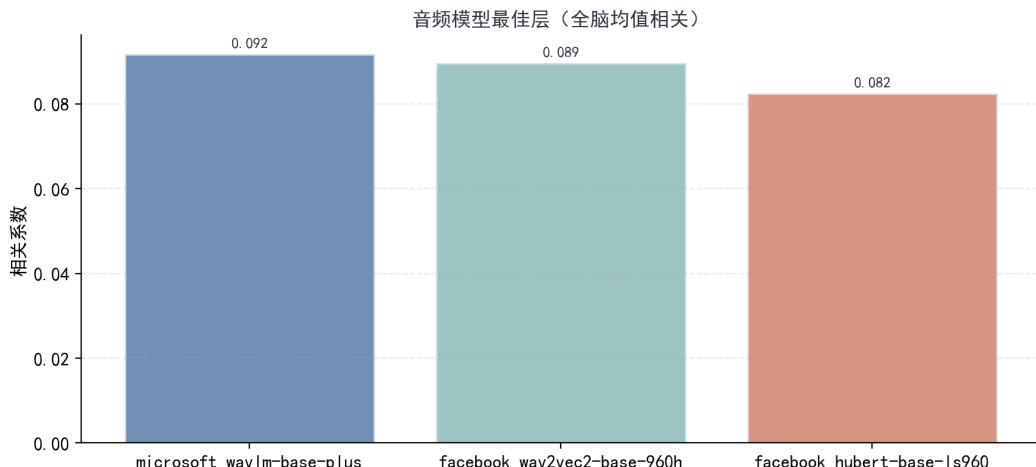


图 6 音频模型最佳层的全脑均值相关系数（由 report/scripts/make_figures.py 从 results/summary.csv 生成）。

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。

7 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

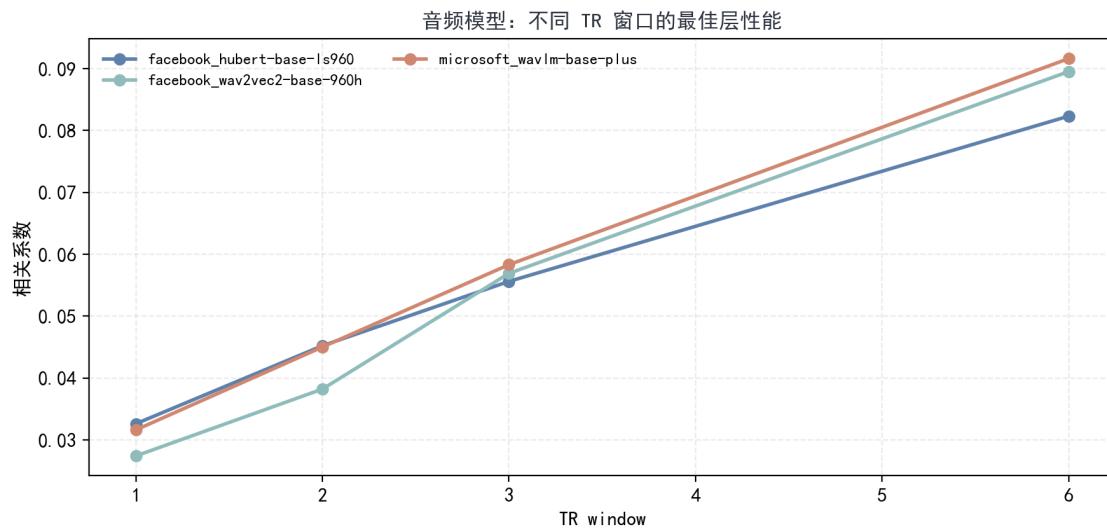


图 7 音频模型在不同 TR 窗口下的最佳层性能趋势（由 `report/scripts/make_figures.py` 从 `results/summary.csv` 聚合生成）。

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。

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。

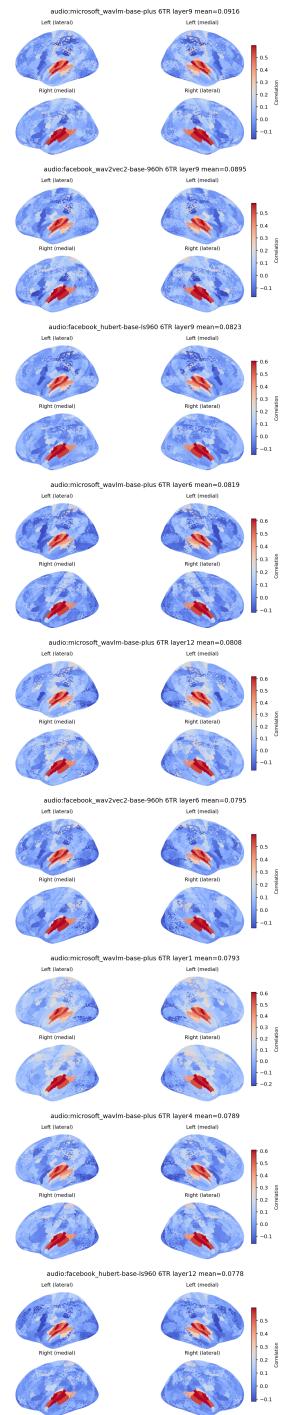


图 8 音频模型类别脑图对照：WavLM、Wav2Vec2、HuBERT 在各自最佳配置下的相关图（由 src/run_plot_corr_maps.py 绘制并组合）。

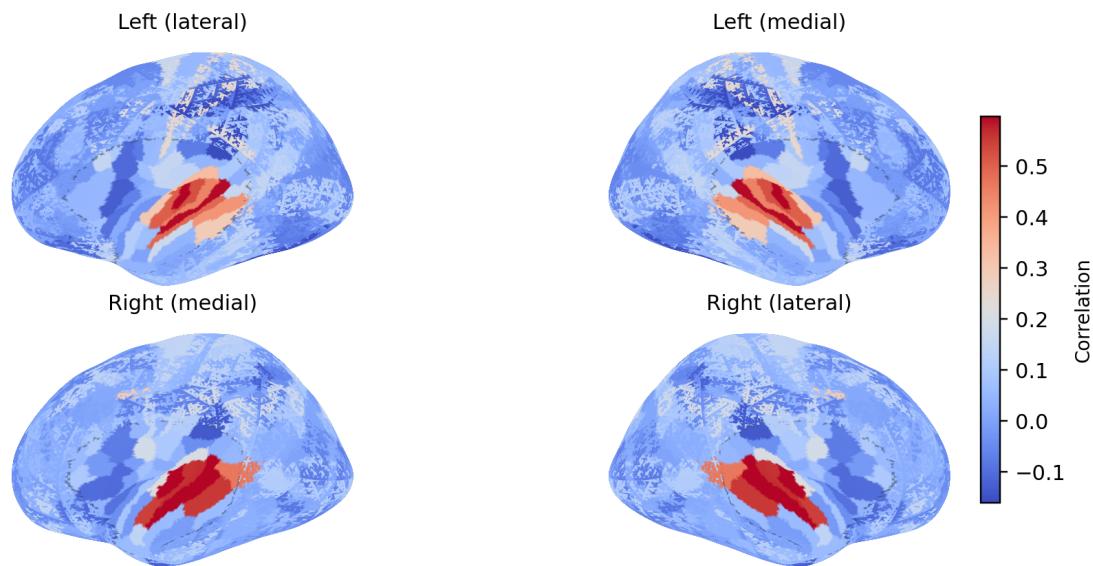


图 9 WavLM-base-plus (6TR, layer9) 相关图可视化（由 `src/run_plot_corr_maps.py` 从 `corr_layer9.npy` 绘制）。

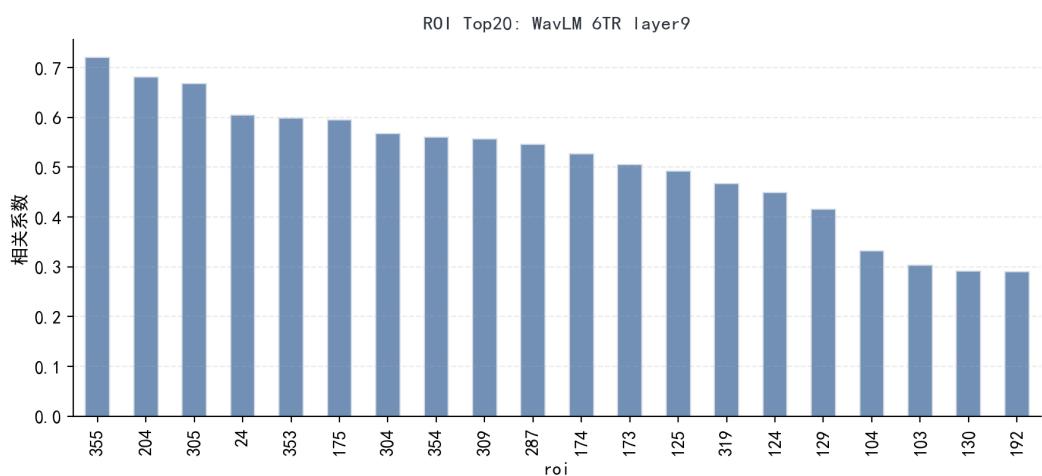


图 10 WavLM-base-plus (6TR, layer9) 对应 corr map 的 ROI Top20 (由 `report/scripts/make_figures.py` 从 `results/roi.csv` 生成)。

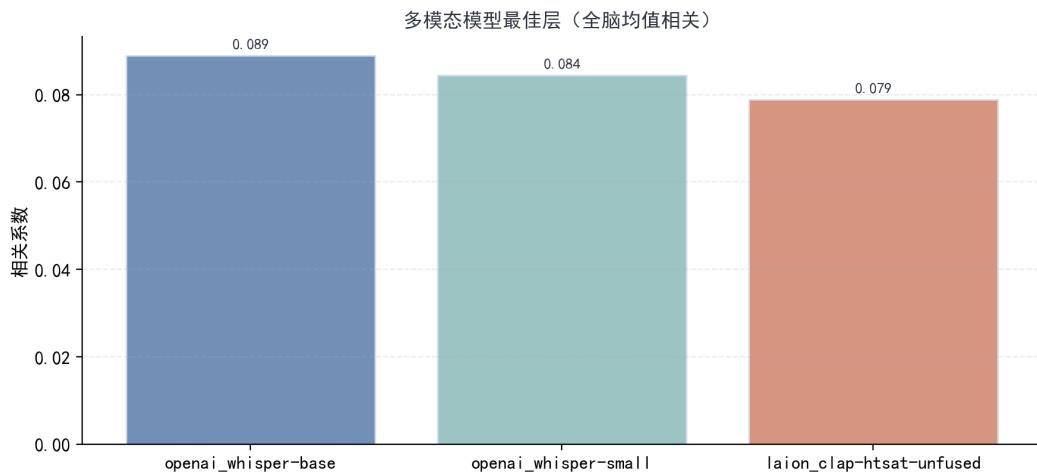


图 11 多模态模型最佳层的全脑均值相关系数（由 report/scripts/make_figures.py 从 results/summary.csv 生成）。

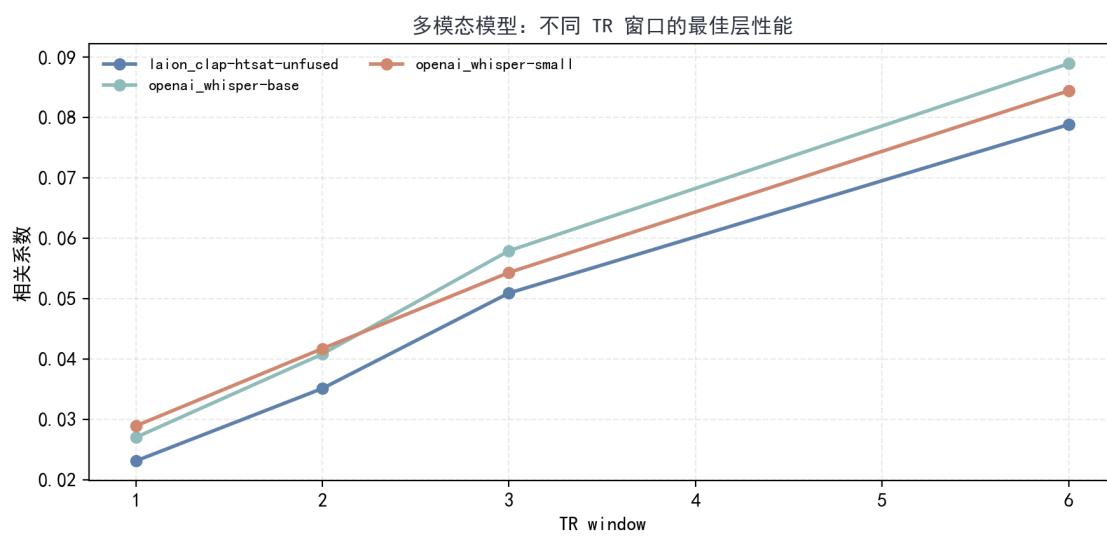


图 12 多模态模型在不同 TR 窗口下的最佳层性能趋势（由 report/scripts/make_figures.py 从 results/summary.csv 聚合生成）。

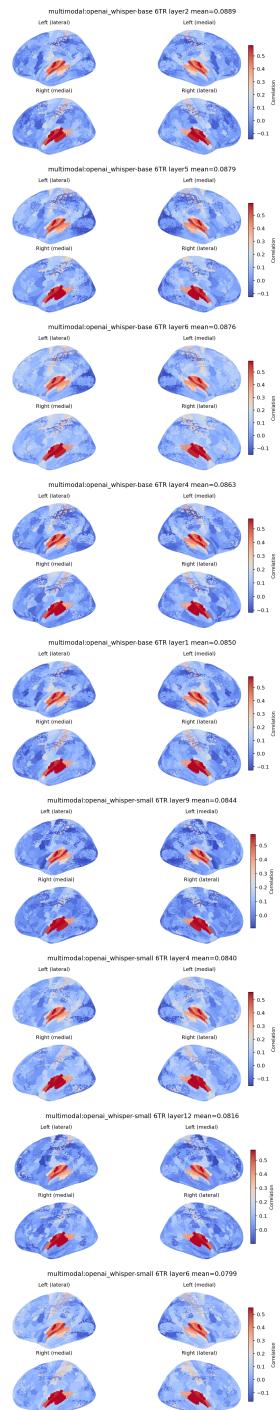


图 13 多模态模型类别脑图对照:Whisper-base(6TR,layer2)与 Whisper-small(6TR,layer9)(由 src/run_plot_corr_maps.py 绘制并组合)。

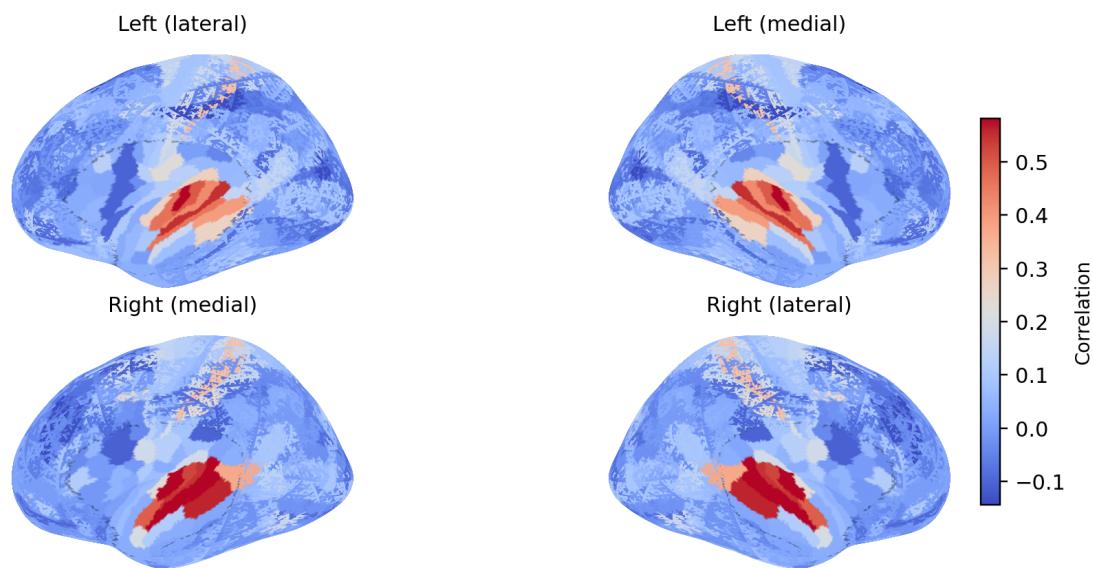


图 14 Whisper-base (6TR, layer2) 相关图可视化 (由 `src/run_plot_corr_maps.py` 从 `corr_layer2.npy` 绘制)。

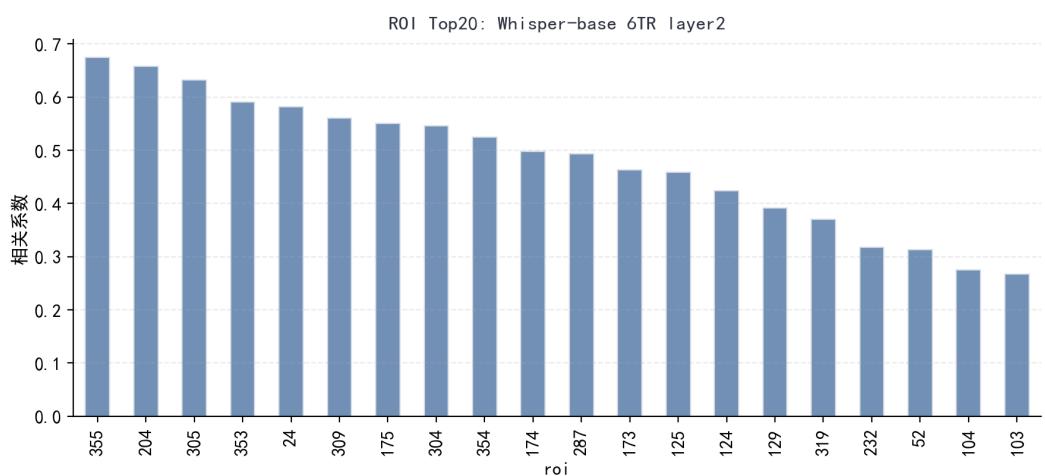


图 15 Whisper-base (6TR, layer2) 对应 corr map 的 ROI Top20 (由 `report/scripts/make_figures.py` 从 `results/roi.csv` 生成)。

8

文本 + 音频融合结果：覆盖范围、最优配置与层交互结构

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。本报告在此仅对当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`

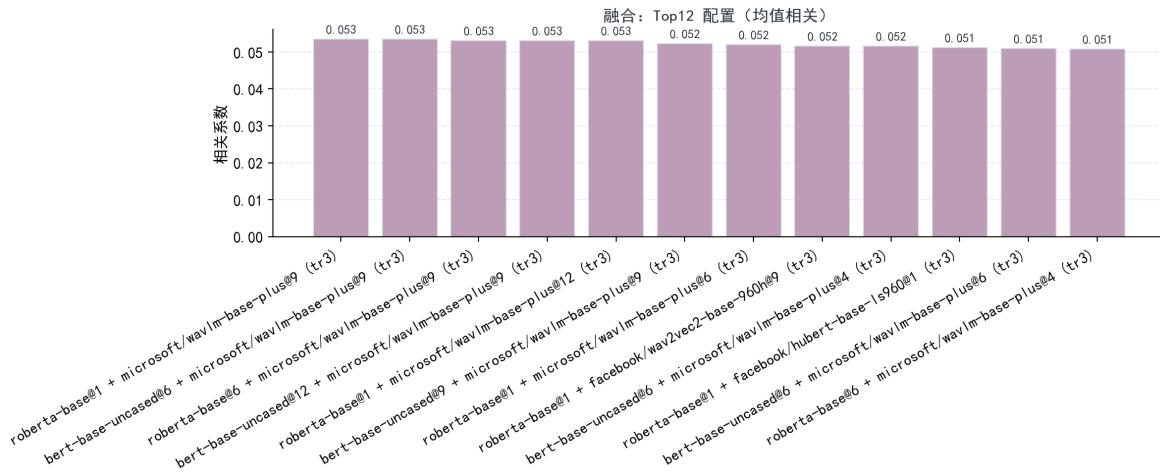


图 16 融合：Top12 配置的全脑均值相关对比（由 report/scripts/make_figures.py 从 results/fusion/**/log.txt 解析生成）。

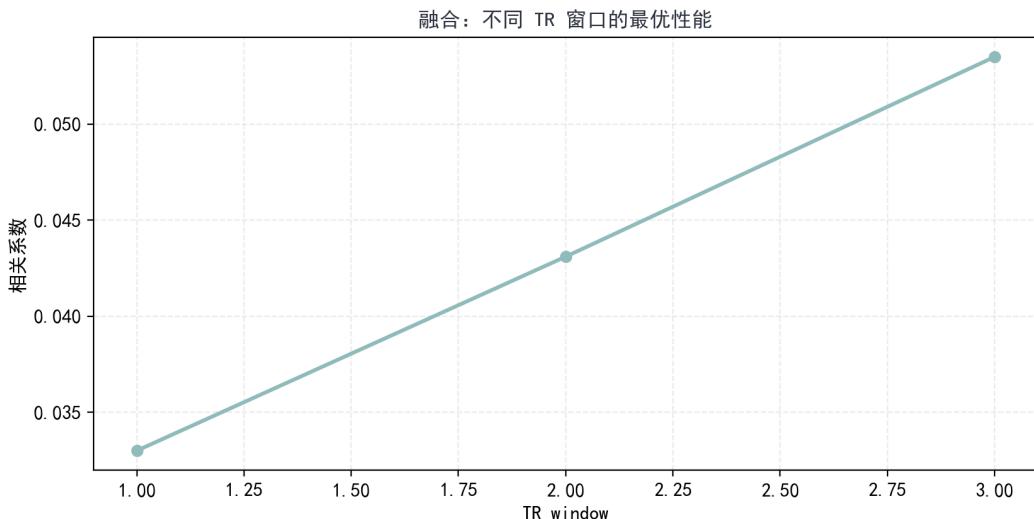


图 17 融合：不同 TR 窗口下的全局最优性能趋势（由 report/scripts/make_figures.py 从融合日志解析生成）。

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，并检查融合是否在部分 ROI 上更接近文本或音频的模式。

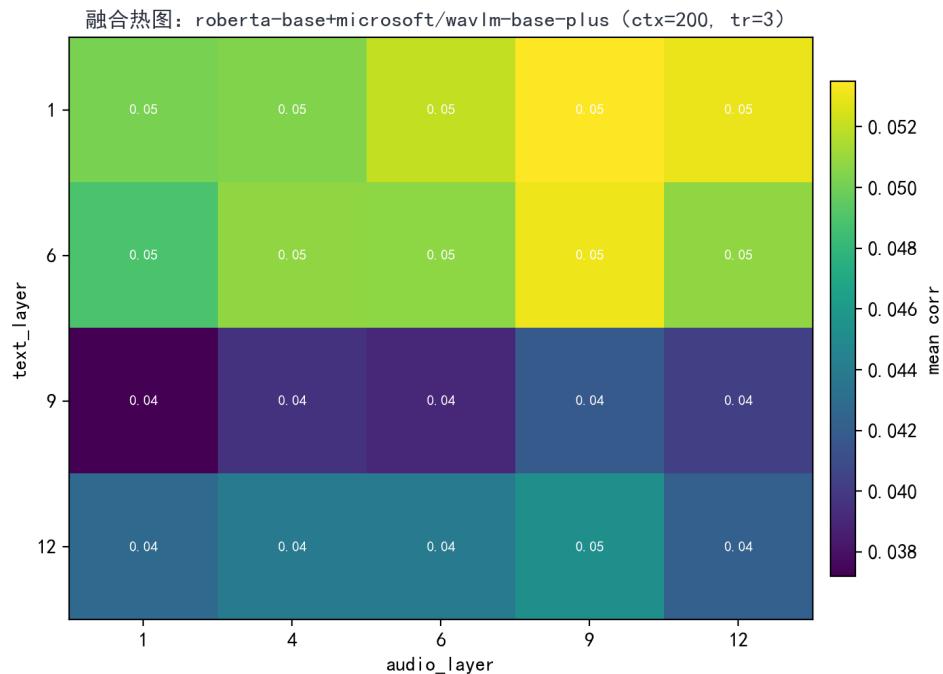


图 18 融合热图：全局最优模型对 (BERT + WavLM) 在 ctx=200、tr=3 条件下的层交互结构 (由 report/scripts/make_figures.py 解析融合日志并绘制)。

9

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

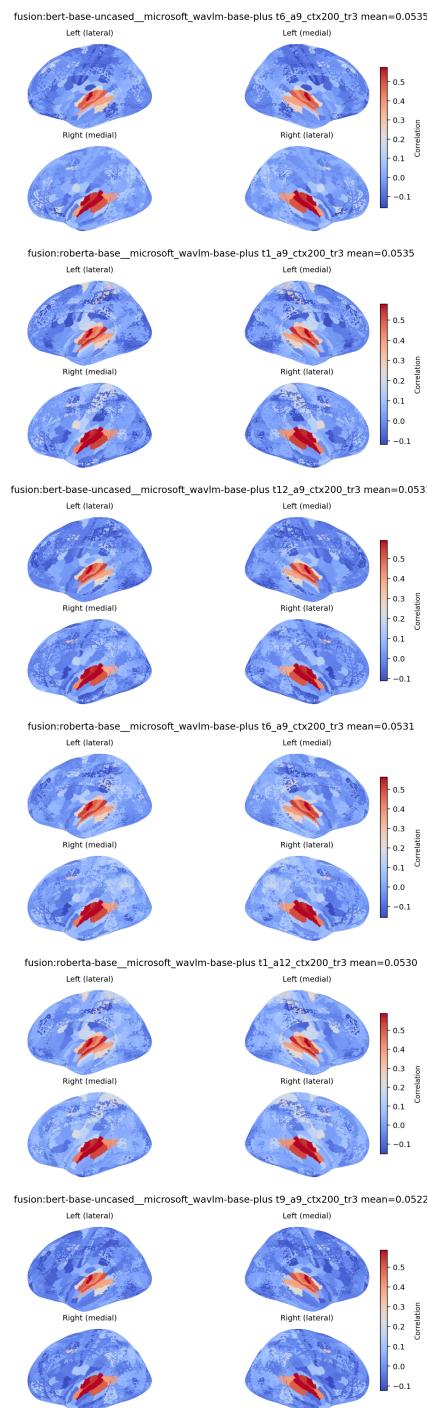


图 19 融合脑图对照：融合 Top6 配置的 corr map 脑图组合（由 src/run_plot_corr_maps.py 从 results/fusion 的 corr_*.npy 选取并绘制）。

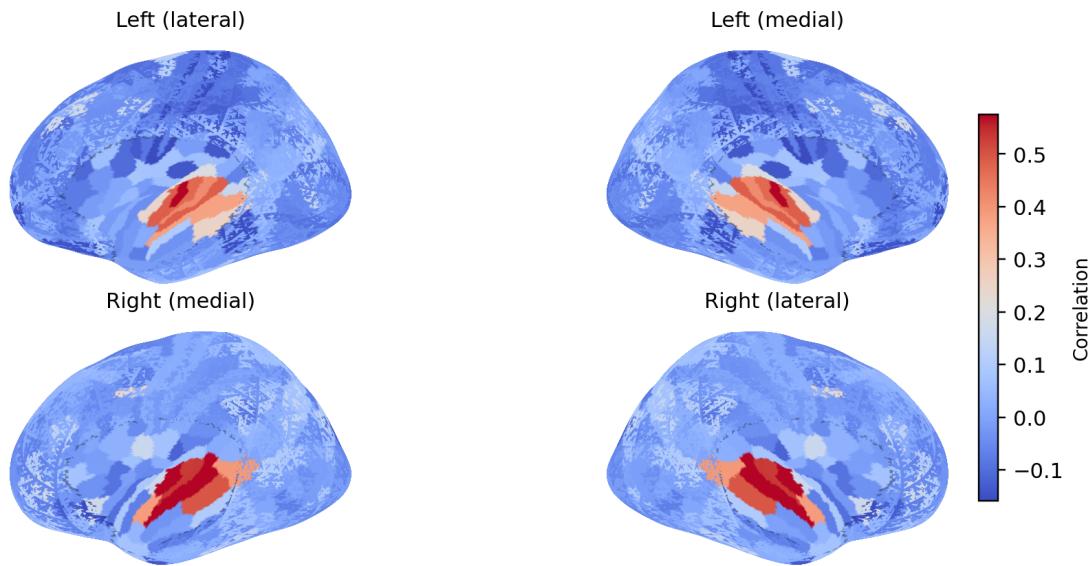


图 20 全局最优融合配置脑图：bert-base-uncased (layer6) + WavLM-base-plus (layer9), ctx=200, tr=3 (由 src/run_plot_corr_maps.py 从对应 corr_*.npy 绘制)。

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。

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。

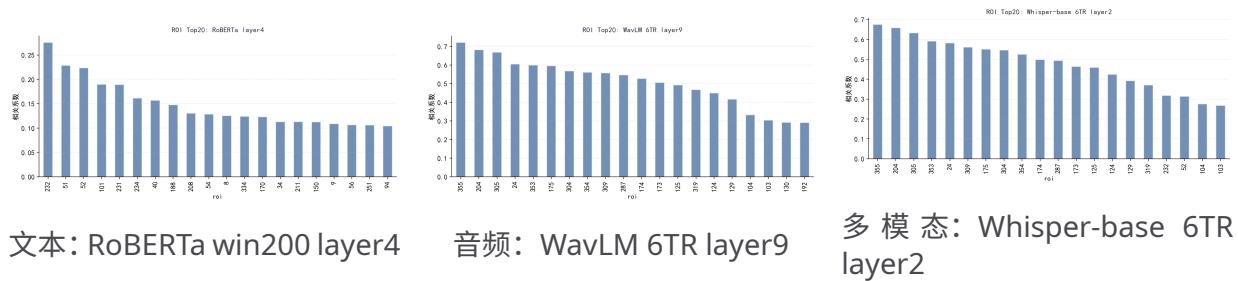


图 21 三种代表性配置的 ROI Top20 对照 (由 report/scripts/make_figures.py 从 results/roi.csv 生成)。

10

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/`。