用于预测对自然刺激的大脑反应的多模态 SEQ2SEQ TRANSFORMER

Qianyi He
Data Science Institute
University of Chicago, USA
heqianyi926@uchicago.edu

Yuan Chang Leong
Department of Psychology
University of Chicago, USA
ycleong@uchicago.edu

ABSTRACT

Algonauts 2025 挑战邀请社区开发能够预测自然场景多模态电影的全脑 fMRI 反应的编码模型。在本次提交中，我们提出了一种序列到序列的 Transformer，能够自回归地从视觉、听觉和语言输入中预测 fMRI 活动。刺激特征通过预训练模型提取，包括 VideoMAE、HuBERT、Qwen 和 BridgeTower。解码器通过双重交叉注意机制整合了先前脑状态、当前刺激和情节级别概要的信息，这些机制既关注从刺激中提取的感知信息，也关注由叙事内容的高级概要提供的叙事信息。我们方法的一个核心创新是使用多模态上下文序列来预测大脑活动序列，使模型能够捕捉到刺激和神经反应中远距离的时间结构。另一个创新是结合了共享编码器和部分特定于个体的解码器，该解码器利用了跨受试者的共同结构，同时考虑了个体的变异性。我们的模型在同分布数据和异分布数据上均表现出色，证明了对时间感知的多模态序列建模在大脑活动预测中的有效性。代码可在 https://github.com/Angelneer926/Algonauts_challenge 找到。

Keywords neural encoding · sequence-to-sequence transformer · naturalistic movies · fMRI

1 介绍

理解人类大脑如何响应自然的、多模态的刺激是认知神经科学中的一个核心目标。功能性神经影像学的最新进展以及大规模、详细注释数据集的发布，使得开发将感觉输入映射到广泛皮层区域的脑活动的预测模型成为可能。Algonauts 2025 挑战赛为评估此类模型提供了一个独特的试验场，邀请各领域的研究人员参与一个为期 7 个月的挑战，以预测观看包括同步视觉、听觉和语言流的电影刺激的受试者的全脑 fMRI 响应 [1] 。

传统的大脑反应建模方法，例如有限冲激响应（FIR）模型或岭回归，通常学习线性映射，从而独立地从最近的刺激历史预测 fMRI 信号的每个时间点。[2, 3] 。虽然这些方法在许多情况下都取得了成功，但它们常常忽视神经反应的动态、自回归特性，并且在整合多模态输入或适应不同受试者的个体差异方面能力有限。受到最近研究的启发，这些研究表明深度神经网络可以有效地建模对自然语言和视觉的皮层反应 [4, 5, 6] ，我们提出了一种基于 Transformer 的序列到序列架构，以更好地捕捉塑造 fMRI 信号随时间变化的时间依赖性和多模态交互。类似于基于神经网络的机器翻译模型，它们将一个语言的单词序列映射到另一个语言的单词序列 [7] ，我们的模型将视听和语言刺激的序列转化为大脑反应的序列，使每次预测都依赖于完整输入序列和先前神经活动的历史。

为了考虑自然刺激的丰富的、层次化的结构，我们从多模态的最先进预训练模型中提取刺激特征：VideoMAE 用于视觉运动的时间动态 [8] ，HuBERT 用于声学特征 [9] ，Qwen 用于语言表征。此外，我们结合了从 BERT 中导出的句子级语义特征 [10] ，这些特征使用训练数据中电视剧集的高级手动总结来源。这样我们就能提供超越瞬时刺激输入的更广泛的叙述背景 [11] 。为了捕捉联合的视觉-语言表征，我们使用 BridgeTower 从刺激中提取跨模态特征 [12] 。

为了应对大脑反应中的个体差异，我们采用了一种混合架构，包含共享编码器和特定于个体的解码器头。这种方法基于直觉：虽然感知表征可能在很大程度上是共享的，但这些表征如何转换为神经活动可以因人而异 [13, 14] 。这种设计通过在汇总的数据上训练一个共享编码器来提高稳健性，使其能够学习到通用的刺激表征。它还通过仅微调轻量级的、特定于个体的解码器头来实现对新参与者的高效适应。最后，我们的模型利用自回归解码与教师强迫以及可学习的 BOS 标记来预测整个 fMRI 反应序列。我们使用均方误差和皮尔逊

相关损失的结合来优化我们的模型，以便更好地与挑战的评估指标对齐。大量实验表明，这种架构产生了稳健的预测，并有效地整合了时间上的多模态信息。

2 相关工作

大多数现有的 fMRI 编码模型遵循一个标准流程：从连续的刺激中提取特征，在一个短时间窗口内将其拼接，使用线性模型预测逐体素的大脑响应，正则化强度通常通过交叉验证进行调节。尽管这种框架已被证明是有效的，但最近的进展大多集中在优化特征选择，而不是改变模型的整体结构。在语言领域，早期的方法使用静态词嵌入来表示独立于上下文的词义 [15]，而最近的研究表明，来自大型语言模型（LLMs）的上下文化表征在大脑活动预测中具有显著更好的表现 [16]。编码性能随着模型规模的增加大约呈对数增长，且更大的模型揭示出越来越多的左侧化激活模式，这与经典的语言神经科学发现一致 [17]。

在语音编码中，模型已经从使用人工设计的声谱图和音素级特征 [18] 发展到使用自监督模型（如 HuBERT 和 WavLM）的表示，这些表示能够更好地捕捉听觉信息的层次结构，并随着模型规模的扩大显示出改进的编码效果 [19]。在视觉领域，编码模型已经从在 ImageNet 上训练的 CNN 演变为自监督的视频 Transformer，例如 VideoMAE，这些模型能够捕捉时间和运动特征 [20]。更近期的方法利用了来自视觉-语言模型（如 CLIP）的表示，尽管证据表明，这些改进主要归因于大规模训练数据，而不是语言监督本身 [21, 22]。为了更明确地整合跨模态的信息，跨模态 Transformer（如 BridgeTower）通过共同注意力融合语言和视觉特征，从而得到更好匹配多个皮层系统的大脑反应的表示 [23]。

然而，尽管在特征选择和模型缩放方面取得了进展，基础的编码流程仍然基本保持不变。尽管一些近期工作探索了架构上的替代方案，比如在视听区域上对经过预训练的听觉网络进行微调以改进预测 [24]，但对能够共同捕捉时间动态、多模态集成和个体差异性的端到端模型的探索仍然有限。在这篇投稿中，我们朝这个方向迈出了一步，引入了一种序列到序列的转换器，将 fMRI 编码框架化为一个自回归生成任务，在一个统一的框架中结合了时间延伸的跨模态注意机制和个体化的输出头。

3 数据集和挑战

Algonauts 2025 竞赛基于 CNeuroMod 数据集的一个子集 [25]，包括来自四位参与者（sub-01, sub-02, sub-03, sub-05）的全脑 fMRI 记录，在他们观看自然多模态电影时获得。训练数据包括约 65 小时的电影刺激，来自《老友记》第一季到第六季的所有剧集和四部故事片（《谍影重重 2》、《隐藏人物》、《生命》和《华尔街之狼》），每部影片都与时间分辨的 fMRI 响应对齐。fMRI 响应经过预处理，并使用 Schaefer 图谱 [26] 被分割成 1,000 个皮层区域，每 1.5 秒采样一次，并与 MNI 空间对齐。

模型性能通过预测的和真实的 fMRI 响应之间的皮尔逊相关性进行评估。最终得分是通过对所有受试者、测试电影和大脑区块的相关值进行平均计算得出的。这个得分是在最终模型选择阶段计算分布外（OOD）电影的。

4 方法

4.1 特征提取

我们从多种模态中提取刺激表征，以全面捕捉每个时间点的外部输入。参见图 1。本节概述了用于获取和增强这些特征的模型和策略。

对于视觉输入，我们使用了 VideoMAE，一种基于大规模视频数据集预训练的掩码自动编码器。VideoMAE 非常适合这一任务，因为它能够对细粒度的时空依赖性进行建模，尤其是运动动态和场景切换。在每个时间点，我们通过对最终（第 12 层）编码器层的标记表示求平均，提取当前帧的 768 维嵌入。然后，我们将其与所有前面帧的平均嵌入拼接，得到一个 1536 维的向量。这种简单而有效的时间平滑提供了一种记忆形式，帮助模型维持对过去视觉上下文的感知，这对于具有连续叙述流的视频尤其有益。

为了对音频信号进行编码，我们使用了 HuBERT，这是一种通过掩蔽预测潜在单元来学习分层语音表示的自监督模型。我们通过连接第 3 层和第 9 层编码器的隐藏状态，在每个时间点提取了一个 1536 维的特征向量，使得能够捕获语音和韵律信息。虽然我们尝试结合过去的上下文（例如，对之前帧进行平均），但我们发现这种时间聚合在该模态下会降低性能。这可能是由于音频相比于视觉信号具有更高的时间分辨率，其中逐帧的精度对于保留短期声学动态至关重要。因此，对于每一帧，我们仅保留帧级特征，而不结合之前的时间点。

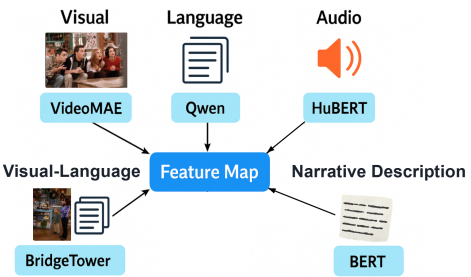

Figure 1: 特征提取

我们使用了 Qwen 模型来嵌入对话记录。在每个时间点，模型接收当前话语及之前所有话语作为文本上下文，并被截断到最近的 2048 个标记。在第 12 层 transformer 层的隐藏状态中，我们计算了两种类型的特征：（1）对完整输入序列中所有标记的平均值，以及（2）仅对最后 10 个标记的平均值。每种特征都是一个 1024 维的向量，它们的连接组合形成了一个 2048 维的表示，该表示同时捕捉了全局语义上下文和当前话语边界附近的局部信息。

为了补充单模态特征，我们使用了 BridgeTower，一种预训练的跨模态变压器架构，旨在整合视觉和文本信号。在每个时间点，该模型接收一个视频帧及其相应的语句作为输入。我们从视觉/文本融合后的 [CLS] 标记对应的最终层池化器输出中提取一个 1536 维的融合表示。这些特征提供了额外的跨模态对齐，这在独立的视觉和语言编码器中难以获得，并且作为特定模态表示的重要补充。

虽然原始数据集提供了对话级别的注释，但其中的情节包含对解读至关重要的总体叙事结构。为了弥补这一缺失的上下文，我们手动从网上资源（例如，Google 搜索）中获取了剧集级别的摘要。每个摘要被分成句子，并通过一个预训练的 BERT 模型来获得句子级别的嵌入。我们探索了两种策略将剧集级别的语言上下文融入帧级表示学习过程。第一种是基于每个帧的时间距离，对句子嵌入进行的高斯加权求和。第二种是在解码器中的交叉注意力机制，其中每个时间步的 fMRI 预测关注来自剧集摘要的句子级别嵌入。尽管高斯加权方法施加了一种平滑且可解释的时间偏差，交叉注意力通过允许模型学习帧与叙事内容之间的动态对齐提供了更大的灵活性。虽然摘要与帧之间的对齐是近似的，这种额外的模态捕捉了高层次的主题信息，可能指导解读，特别是在长篇剧集中。

4.2 模型架构

我们提出了一种基于 Transformer 的编码器-解码器架构，用于建模从多模态刺激特征到 fMRI 反应的时间映射。该设计考虑了神经活动的序列动态和被试间的变异性。与传统的静态回归方法不同，我们的模型以自回归方式生成 fMRI 时间序列，采用掩码自注意力和跨模态上下文集成自低级刺激和高级语义表示。整体框架如图 2 所示。

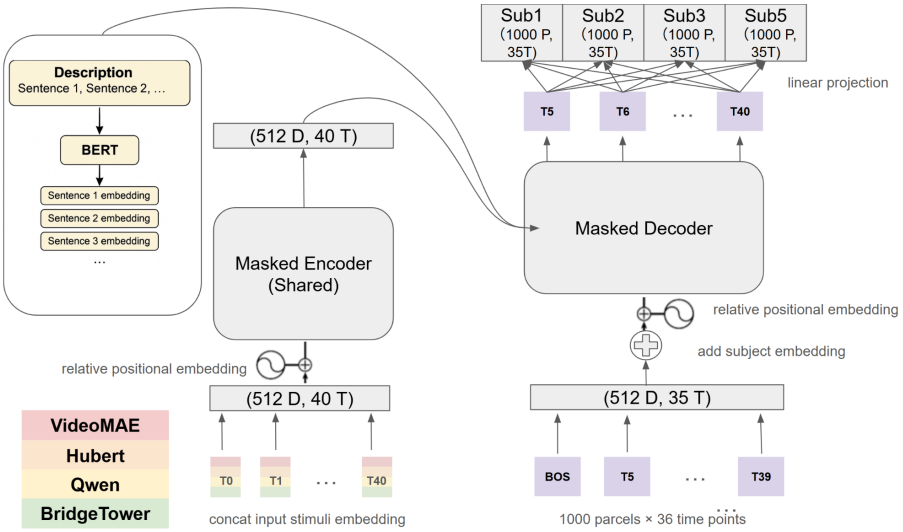

Figure 2: 模型架构

编码器被实现为具有因果自注意力的 Transformer，这样每个时间步仅关注当前和之前的输入。这不同于机器翻译中使用的 seq2seq Transformers 的标准双向设计，其动机来自神经处理的本质：大脑按照时间展开的顺序来编码外部刺激，而无需访问未来信息。通过施加这种时间约束，模型可以更好地与生物感知的动态对齐。实证结果表明，与不受限制注意力的编码器相比，这种修改带来了更好的预测性能。在每个编码器层之前，我们为输入序列添加了相对位置编码，使模型可以以位置不变的方式处理时间。令 $\boldsymbol{H}_{\text{enc}} = \text{Encoder}(\boldsymbol{X}) \in \mathbb{R}^{T \times d}$ 表示编码器的输出，其中 T 是时间步数，d 是隐藏维度。

解码器是一个掩码因果 Transformer，它以自回归的方式生成 fMRI 信号。让 $\hat{\boldsymbol{Y}} = [\hat{\boldsymbol{y}}_1, \hat{\boldsymbol{y}}_2, \ldots, \hat{\boldsymbol{y}}_T]$ 表示预测的 fMRI 序列。在每个解码步骤 t 中，解码器将先前生成的输出 $\hat{\boldsymbol{y}}_{t-1}$ 的嵌入作为查询，并仅通过掩码自注意

力关注过去，确保在时间 t 的预测仅依赖于在 $t' \leq t$ 时可用的信息。这个约束模拟了受试者观察刺激的时间单向性。

为了给解码器提供高层次的语义上下文，我们引入了一个额外的跨注意力模块，该模块应用于通过冻结的 BERT 模型提取的情节级描述 $\boldsymbol{E}$ 。该语义记忆在对刺激特征进行标准跨注意力处理之后被整合，使解码器在 fMRI 预测过程中能够同时关注感知信息和叙述信息。

形式上，解码器在时间 t 的输出计算为：

$$
\begin{aligned}
\tilde{\boldsymbol{y}}_t &= \hat{\boldsymbol{y}}_{t-1}\boldsymbol{W}_{\text{dec}} + \boldsymbol{b}_{\text{dec}} \\
\boldsymbol{z}_t^{(0)} &= \tilde{\boldsymbol{y}}_t + \text{RelPosEnc}(t) \\
\boldsymbol{z}_t^{(l,1)} &= \text{LayerNorm}\left(\boldsymbol{z}_t^{(l-1)} + \text{SelfAttn}(\boldsymbol{z}_{\leq t}^{(l-1)})\right) \\
\boldsymbol{z}_t^{(l,2)} &= \text{LayerNorm}\left(\boldsymbol{z}_t^{(l,1)} + \text{CrossAttn}_{\text{stim}}(\boldsymbol{z}_t^{(l,1)}, \boldsymbol{H}_{\text{enc}})\right) \\
\boldsymbol{z}_t^{(l,3)} &= \text{LayerNorm}\left(\boldsymbol{z}_t^{(l,2)} + \text{CrossAttn}_{\text{desc}}(\boldsymbol{z}_t^{(l,2)}, \boldsymbol{E})\right) \\
\boldsymbol{z}_t^{(l)} &= \text{LayerNorm}\left(\boldsymbol{z}_t^{(l,3)} + \text{MLP}(\boldsymbol{z}_t^{(l,3)})\right) \\
\hat{\boldsymbol{y}}_t &= \boldsymbol{z}_t^{(L)}\boldsymbol{W}_{\text{out}} + \boldsymbol{b}_{\text{out}}
\end{aligned}
\tag{1}
$$

。所有注意力模块都是多头的，并包含 dropout 和残差连接。

训练目标结合了均方误差（MSE）和预测序列与真实值序列之间的负皮尔逊相关性：

$$
\mathcal{L}_{\text{MSE}} = \frac{1}{T}\sum_{t=1}^{T}\|\hat{\boldsymbol{y}}_t - \boldsymbol{y}_t\|_2^2, \quad \mathcal{L}_{\text{corr}} = -\frac{1}{T}\sum_{t=1}^{T}\rho(\hat{\boldsymbol{y}}_t, \boldsymbol{y}_t), \quad \mathcal{L} = \mathcal{L}_{\text{MSE}} + \lambda\mathcal{L}_{\text{corr}} \tag{2}
$$

其中，ρ 表示皮尔逊相关系数，λ 是一个加权因子。

为了支持用相同的模型对多个受试者进行建模，我们引入了一种混合架构，其中编码器在所有受试者之间共享，而解码器则结合了个别特定的组件。这个设计反映了即使在相同刺激下，fMRI 响应在受试者之间也表现出显著的差异性。同时，利用来自多个受试者的数据可以提高模型的稳健性和泛化能力。每个受试者 s 都关联有一个可学习的嵌入向量 $\boldsymbol{e}_s$ ，这个向量被连接到解码器输入的每个时间步。此外，解码器头部的最终线性投影层是针对每个受试者专门设计的，这允许从隐藏表示到预测 fMRI 输出的个性化映射。这个设置假设刺激如何被表示可以在个体之间共享，而相应的脑响应则依赖于个体。形式上，对于受试者 s ，对原始架构进行了两处修改：

$$
\begin{aligned}
\boldsymbol{z}_t^{(0)} &= \tilde{\boldsymbol{y}}_t + \boldsymbol{e}_s + \text{RelPosEnc}(t) \\
\hat{\boldsymbol{y}}_t^{(s)} &= \boldsymbol{z}_t^{(L)}\boldsymbol{W}_{\text{out}}^{(s)} + \boldsymbol{b}_{\text{out}}^{(s)} \quad \text{(for subject } s\text{)}
\end{aligned}
\tag{3}
$$

值得注意的是，这种架构通过仅微调特定于主体的嵌入和输出头，实现了对新主体的参数高效适应，使其适合于 fMRI 编码任务中的小样本个性化。

为了解决在脑反应序列级别预测中训练数据有限的挑战，我们使用滑动窗口的方法扩充了数据集。具体来说，我们将原始刺激分割成重叠的时间块，每个块由 40 个连续的时间步骤的多模态输入组成。相应的输出是一个 35 帧的 fMRI 序列，在应用固定血流动力学延迟后与输入窗口在时间上对齐。这种策略增加了有效的训练样本数量，并确保每个 fMRI 时间点在不同的重叠窗口中被多次预测，从而实现更稳定和强健的学习。

为了减轻由于上下文重复导致的过拟合，我们在每个训练周期开始时打乱了训练样本的顺序。这防止模型记忆固定的刺激-响应映射，并鼓励其学习不同时间上下文中的可泛化特征。窗口长度和血液动力学延迟都被视为可调节的超参数。虽然当前设置表现出很强的实际性能，但全面的超参数搜索仍是未来的一个研究方向。

在训练过程中，我们用一个可学习的 `[BOS]` 标记初始化了解码器，并采用教师强制策略来指导自回归生成。在每个解码步骤 t ，模型接收地面事实输出 $\boldsymbol{y}_{t-1}$ 或其自身先前的预测 $\hat{\boldsymbol{y}}_{t-1}$ 作为输入，依据教师强制比例 $\gamma \in [0, 1]$ 进行选择。以概率 γ ，使用地面事实以稳定训练和加速收敛；以概率 $1-\gamma$ ，模型从其自身的输出分布中进行采样，使其暴露于潜在的预测错误中，并在推理时提高鲁棒性。教师强制比例在训练历元中逐渐下降，从而逐步从监督指导过渡到自条件生成。

我们使用包含分布内（ID）和分布外（OOD）数据的验证集和测试集来评估我们的模型。

为了评估模型开发过程中的性能，我们构建了一个验证集，其中包括《老友记》第六季的所有 b 部分集，以及每部 Movie10 电影的前两个片段（《谍影重重》、《隐藏人物》、《生命》和《华尔街之狼》）。这一划分包括

《老友记》和电影内容的混合，能够可靠地评估不同类型内容。在这个验证集上，模型在未使用的《老友记》集上实现了平均 Pearson 相关系数为 0.301，在未使用的 Movie10 片段上实现了相关系数为 0.225。详见图 3 。

朋友集数据与电影 10 的性能差距可能是由于训练集中包含了更大比例的朋友数据，从而使模型能够更有效地学习该刺激集中的模式。相比之下，电影 10 片段在视觉风格和叙事结构上有很大不同，对模型的泛化性提出了更大的挑战。因此，电影 10 的较低表现是可以预期的，反映了从单一系列的情节电视节目中学习的表示转移到多样化电影特征集的挑战。另一个可能的原因是缺乏电影的剧集级摘要，这限制了模型在朋友剧集中可以利用的高级语义上下文的获取。

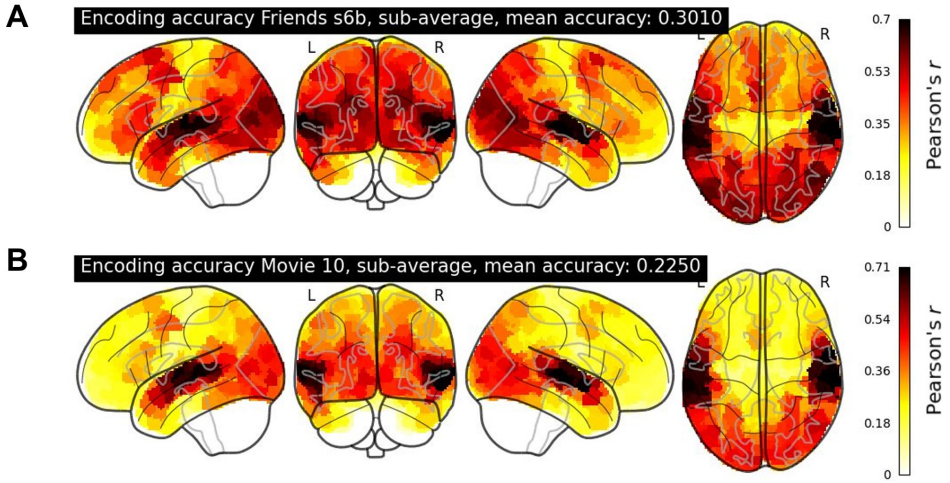

Figure 3: (A)《老友记》第 6 季 b 的模态验证准确性 (B) 剪切出来的电影片段

然后，我们在 Algonauts 2025 挑战赛提供的官方测试数据上评估了我们的最终模型，该数据分为两个阶段。在模型构建阶段，测试性能是在《老友记》第七季的剧集中测量的，代表一种分布内（ID）评估设置。在随后的模型选择阶段，模型在一组未见过的长篇电影上进行评估，这构成了一个真正的分布外（OOD）泛化任务。

我们的模型在《老友记》第七季上实现了 0.305 的平均皮尔逊相关系数，而在 OOD 电影集上达到了 0.199。参见图 4 。第七季上的强劲表现表明模型能够在其主要训练的领域内可靠地泛化。与我们的验证结果相似，电影上的表现显著较低，这表明模型在向分布外刺激转移时效果不如预期。

在这项研究中，我们提出了一种多模态序列到序列的 Transformer 模型，用于预测对自然视听和语言刺激的全脑 fMRI 响应。通过利用跨多种模态预训练的编码器、结合来自剧集摘要的语义层次上下文以及使用特定于受试者的自回归适应来建模脑活动，我们的方法捕捉到了神经响应的时间复杂性和个体复杂性。该模型在分布内和分布外数据上表现出强劲性能，突显了时间感知和跨模态整合的优势。此外，结合教师强制的滑动窗口训练策略，即使数据有限，也能实现高效学习。未来方向包括结合更精确的高级语义上下文与神经动力学之间的对齐、将模型扩展到更广泛的受试者群体，以及探索跨未见个体和刺激的零样本泛化。我们的结果支持生成性多模态模型作为在现实世界环境中连接感知与脑活动的强大框架的潜力。

5

致谢 本研究利用了芝加哥大学数据科学研究所的高性能计算集群提供的资源完成。

References

[1] Alessandro T. Gifford, Domenic Bersch, Marie St-Laurent, Basile Pinsard, Julie Boyle, Lune Bellec, Aude Oliva, Gemma Roig, and Radoslaw M. Cichy. The algonauts project 2025 challenge: How the human brain makes sense of multimodal movies, 2025.

[2] Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fmri. *Neuroimage*, 56(2):400–410, 2011.

[3] Tom Dupré la Tour, Matteo Visconti di Oleggio Castello, and Jack L Gallant. The voxelwise encoding model framework: a tutorial introduction to fitting encoding models to fmri data. *Imaging Neuroscience*, 3:imag_a_00575, 2025.

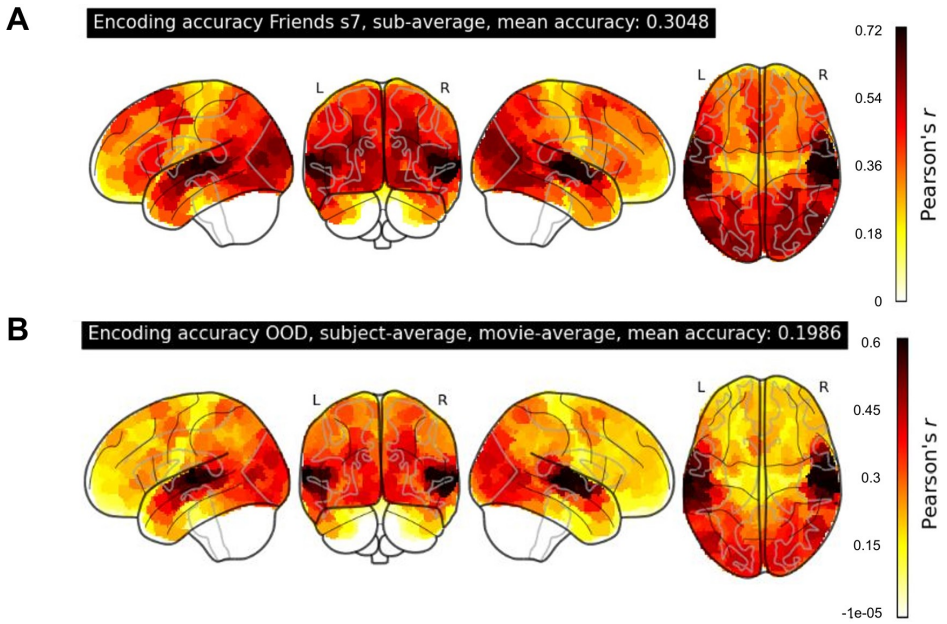

Figure 4: 模态测试精度：(A) 模型构建阶段（《老友记》第七季）(B) 模型选择阶段（分布外电影）

[4] Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications biology*, 5(1):134, 2022.

[5] Martin Schrimpf, Idan Asher 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):e2105646118, 2021.

[6] Haibao Wang, Lijie Huang, Changde Du, Dan Li, Bo Wang, and Huiguang He. Neural encoding for human visual cortex with deep neural networks learning "what" and "where". *IEEE Transactions on Cognitive and Developmental Systems*, 13(4):827–840, 2020.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[8] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

[9] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

[10] 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, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[11] Shan Gao, Ryleigh Nash, Shannon Burns, and Yuan Chang Leong. Predicting whole-brain neural dynamics from prefrontal cortex functional near-infrared spectroscopy signal during movie-watching. *Social cognitive and affective neuroscience*, 20(1):nsaf043, 2025.

[12] Xiao Xu, Chenfei Wu, Shachar Rosenman, Vasudev Lal, Wanxiang Che, and Nan Duan. Bridgetower: Building bridges between encoders in vision-language representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10637–10647, 2023.

[13] Ma Feilong, Samuel A Nastase, Guo Jiahui, Yaroslav O Halchenko, M Ida Gobbini, and James V Haxby. The individualized neural tuning model: Precise and generalizable cartography of functional architecture in individual brains. *Imaging Neuroscience*, 1:1–34, 2023.

[14] Ido Tavor, O Parker Jones, Rogier B Mars, SM Smith, TE Behrens, and Saad Jbabdi. Task-free mri predicts individual differences in brain activity during task performance. *Science*, 352(6282):216–220, 2016.

[15] Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS one*, 9(11):e112575, 2014.

[16] Shailee Jain and Alexander Huth. Incorporating context into language encoding models for fmri. *Advances in neural information processing systems*, 31, 2018.

[17] Laurent Bonnasse-Gahot and Christophe Pallier. fmri predictors based on language models of increasing complexity recover brain left lateralization. *Advances in Neural Information Processing Systems*, 37:125231–125263, 2024.

[18] Xue L Gong, Alexander G Huth, Fatma Deniz, Keith Johnson, Jack L Gallant, and Frédéric E Theunissen. Phonemic segmentation of narrative speech in human cerebral cortex. *Nature communications*, 14(1):4309, 2023.

[19] Richard Antonello, Aditya Vaidya, and Alexander Huth. Scaling laws for language encoding models in fmri. *Advances in Neural Information Processing Systems*, 36:21895–21907, 2023.

[20] Jacob Yeung, Andrew F Luo, Gabriel Sarch, Margaret M Henderson, Deva Ramanan, and Michael J Tarr. Reanimating images using neural representations of dynamic stimuli. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5331–5343, 2025.

[21] Guoyuan Yang, Mufan Xue, Ziming Mao, Haofang Zheng, Jia Xu, Dabin Sheng, Ruotian Sun, Ruoqi Yang, and Xuesong Li. Clip-msm: A multi-semantic mapping brain representation for human high-level visual cortex. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 9184–9192, 2025.

[22] Colin Conwell, Jacob S Prince, Kendrick N Kay, George A Alvarez, and Talia Konkle. A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature communications*, 15(1):9383, 2024.

[23] Jerry Tang, Meng Du, Vy Vo, Vasudev Lal, and Alexander Huth. Brain encoding models based on multimodal transformers can transfer across language and vision. *Advances in Neural Information Processing Systems*, 36:29654–29666, 2023.

[24] Maelle Freteault, Maximilien Le Clei, Loic Tetrel, Lune Bellec, and Nicolas Farrugia. Alignment of auditory artificial networks with massive individual fmri brain data leads to generalisable improvements in brain encoding and downstream tasks. *Imaging Neuroscience*, 3:imag_a_00525, 2025.

[25] Julie Boyle, Basile Pinsard, Valentina Borghesani, Francois Paugam, Elizabeth DuPre, and Pierre Bellec. The courtois neuromod project: quality assessment of the initial data release (2020). In *2023 Conference on Cognitive Computational Neuroscience*, pages 2023–1602, 2023.

[26] Alexander Schaefer, Ru Kong, Evan M Gordon, Timothy O Laumann, Xi-Nian Zuo, Avram J Holmes, Simon B Eickhoff, and BT Thomas Yeo. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri. *Cerebral cortex*, 28(9):3095–3114, 2018.