大型语言模型的隐藏状态可改进脑电图表征学习与视觉解码

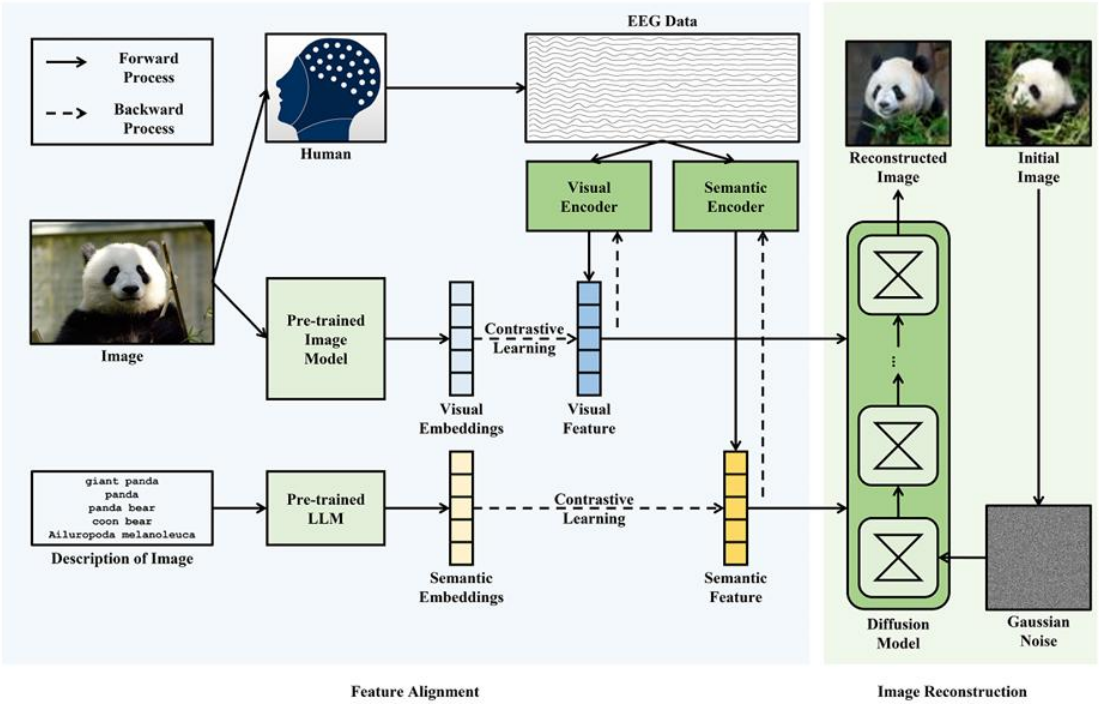
1．选题理由

脑电图（EEG）作为一种非侵入性神经生理技术，可记录大脑神经活动，已吸引了众多研究人员的关注。通过捕捉脑电信号的电位变化，EEG 能提供高时间分辨率的大脑活动信息，揭示与各种神经生理过程相关的动态特性。这些特点使 EEG 成为研究神经活动与认知功能关系的重要工具。近年来，许多研究人员已成功利用深度神经网络从脑信号表征中解码视觉信息。然而，这些研究均面临着共同的挑战与局限：首先，EEG 信号的噪声干扰和稀疏性限制了信息提取的精度；其次，现有方法在复杂语境下对语义信息的解码能力相对有限，尤其是在抽象概念表征方面；此外，对 EEG 信号中语义信息的动态变化缺乏全面理解，限制了方法对不同任务和认知状态的适应性。

大型语言模型（LLMs）的发展可能为解决 EEG 语义信息提取难题提供帮助。目前已有相关研究致力于利用预训练 LLM（如生成式预训练 Transformer（GPT）系列、BERT 等）从文本中提取语义信息，并将其转化用于其他任务。这些模型能够捕捉词语、短语和句子之间的关联，在文本层面编码丰富的语义信息，且该信息可迁移至下游任务。因此，作为文本信息的有效编码结果，LLM 的语义表征或许可进一步应用于脑-文本生成、脑-语义识别等 EEG 语义解码任务。

然而，除语义信息提取外，该领域的另一大挑战是如何从视觉信息中重建刺激并提升重建性能。研究人员尝试将获取的语义信息与生成模型相结合，生成与文本描述匹配的图像，确保生成图像与所提供的语义特征保持一致。这类方法的优势在于无需明确提供图像样本，即可通过端到端的方式从文本生成图像，为基于语义信息的图像生成任务提供了新路径。

为提升 EEG 信号的解码性能，本研究提出一种将 LLM 引入 EEG 视觉解码和刺激图像重建任务的方法。基于 LLM 与人类认知表征的相似性，我们利用 LLM 的语境理解能力生成特征表征，将其作为 EEG 的目标语义特征，从而从 EEG 信号中提取语义特征。具体而言，我们通过向预训练 LLaMa-2 模型输入图像文本描述，获取其隐藏状态的表征，随后通过对比学习使 EEG 表征与 LLaMa-2 嵌入空间对齐；同时从 EEG 信号中提取视觉特征，再利用预训练扩散模型生成图像。通过从训练数据集中筛选图像，并将 EEG 语义特征映射到文本嵌入空间，我们有效融合了两种模态的特征，最终重建出在语义和视觉特征上均与刺激图像相似的图像。实验结果表明，该方法在语义分类准确率和图像重建指标的定量与定性评估中均达到了最优水平。方法整体框架如图 1 所示。



本研究提出的创新方法将语言模型与脑活动数据相结合，克服了以往研究中的部分限制，为更准确、全面地解读 EEG 信号中的信息提供了新视角。总体而言，预训练 LLM 的融入有效增强了对认知信息的分析和拟合能力，为通用人工智能的进一步发展提供了支持。

2．贡献与创新点

2.1 语言-图像预训练

CLIP 首次提出于图像-文本配对任务，旨在通过对比学习学习图像与其文本描述之间的关联。对于多模态特征对齐，CLIP 是一种高效的方法，已被用于脑解码任务中脑活动与刺激特征的对齐。本研究中，采用 CLIP 提出的损失函数（记为 CLIP 损失）。给定一批来自两种模态的数据嵌入 $A=\{a_i \mid i=0,...,M\}$ 和 $B=\{b_i \mid i=0,...,M\}$，对比损失定义为：

$$Con(A,B;\tau)=-\frac{1}{M}\sum_{i=0}^{M}\log[\frac{\exp(\cos(a_i,b_i)/\tau)}{\sum_{j=0}^{M}\exp(\cos a_i,b_j/\tau)}]$$

其中，M 为批次大小，τ 为温度超参数，(A_i,B_i). 表示两种模态数据的匹配嵌入对，$(A_i,B_j)_{j\neq i}$ 其余表匹配对。为在两种模态之间进行双向对比学习，CLIP 损失如下：

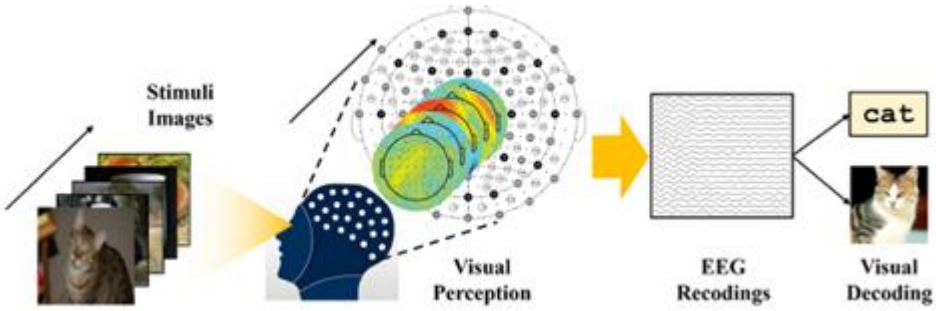
$$L_C(A,B)=\frac{1}{2}(Con(A,B;\tau)+Con(B,A;\tau))$$

通过最小化 CLIP 损失，模型将在共同嵌入空间中，使来自不同模态但含义匹配的两个样本在余弦度量上学习到相似的表征，从而从原始数据中获得高效的表征。预训练模型（LLaMa-2 和 VGG-19）：本任务旨在利用大型语言模型处理图像描述的语义信息。LLaMa-2 模型有三种不同规模：70 亿参数（7b）、130 亿参数（13b）和 700 亿参数（70b）。我们选择了 LLaMa-2 7b，该模型运算速度快，适用于摘要、分类等基础任务。

不同的图像潜在表征会影响 EEG 对视觉刺激的解码效率。在监督任务上预训练的图像模型可能会取得较好的性能。例如，视觉几何组（VGG）预训练模型是一类卷积神经网络（CNN）架构，广泛应用于计算机视觉任务（尤其是图像分类）。其架构简洁统一，通过堆叠多个卷积层和最大池化层，使模型在信息传递过程中能够从输入图像中捕捉到越来越复杂的特征。为更高效地提取图像的视觉特征，我们选择了预训练 VGG-19 模型，该模型已被证实在脑视觉解码的图像检索任务中表现出高性能。

2.2 任务定义

本研究的任务是从受试者观看一组图像时记录的脑活动中解码视觉刺激。设（EEG，图像）对的数据集为$\Omega=(X_i,Y_i),X_i\in S^{C\times T}$为记录的 EEG 信号片段，$Y_i\in I^{H\times W\times 3}$. 为同时呈现的图像，C 为 EEG 信号的通道数，T 为一个时间窗口内的时间点数。EEG 数据在受试者观看刺激图像时记录。本研究的目标是从 EEG 信号片段中生成图像，力求生成的图像在高层和低层特征上均与真实图像高度相似。EEG 视觉解码任务的示意图如下图所示。



2.3 多模态特征提取

低层视觉特征（颜色、形状、结构）和高层语义特征（类别、内容）均可在脑活动中引发相应的模式。因此，从原始 EEG 信号中提取这两个层次的特征进行图像重建是可行的。本研究通过训练深度神经网络，实现特征的自动学习和提取。

EEG 编码器网络需要处理多尺度的时空信息。EEGNet 是一种紧凑的卷积神经网络，包含一个卷积块、一个深度卷积块和一个可分离卷积块，在不指定任务范式的情况下，在许多 EEG 分类任务中都表现出了较高的准确率。与普通卷积层相比，深度卷积块和可分离卷积

块的参数更少，这有助于模型学习更高效的表征并实现更高的泛化能力。我们对 EEGNet 的网络结构进行了修改，在时间卷积块和深度卷积块之间添加了一个卷积块，以整合来自不同通道的信息。该架构分别应用于语义编码器和视觉编码器，记为 E_S and E_V 。

输出特征维度设置为与从预训练模型中获得的目标表征维度相同。训练后，编码器网络将 EEG 信号投影到目标嵌入空间，并从 EEG 信号中预测相应的特征。网络中卷积层的设置细节如表 1 所示。

表 1. 两个编码器网络中四个卷积层的设置细节

Location		Channels	Kernel Size	Padding	Groups
Semantic Encoder Network	1	$1 \rightarrow 64$	(1,41)	(0,20)	1
	2	$64 \rightarrow 2048$	(128,1)	0	64
	3	$2048 \rightarrow 2048$	(1,15)	(0,7)	2048
	4	$2048 \rightarrow 64$	(1,1)	0	1
Visual Encoder Network	1	$1 \rightarrow 16$	(1,41)	(0,20)	1
	2	$16 \rightarrow 64$	(14,1)	0	1
	3	$64 \rightarrow 64$	(1,15)	(0,7)	64
	4	$64 \rightarrow 16$	(1,1)	0	1

我们在训练数据集上训练这两个网络，并根据验证数据集选择最终的编码器轮次。

2.2.1 语义特征

作为语言模型，当向 LLaMa-2 输入文本时，模型会被激活，其隐藏状态能够反映输入文本的语义特征。以往研究表明，从语言模型的中间层可以提取到最佳的语义特征。在 LLaMa-2 7b 模型的 33 层中，我们选择了第 20 层，并记录该层的隐藏状态。对于每个刺激图像，我们将 ImageNet 数据集的描述输入模型，并将隐藏状态在第一维度上的平均值作为图像刺激的语义表征。

LLaMa 模型中隐藏层的宽度为 4096。为提高学习到的语义表征的效率，我们通过一个全连接网络将隐藏层输出的平均值映射到 160 维嵌入空间。相应地，我们也将 EEG 语义编码器的输出维度设置为 160。设图像的文本描述为 t，图像的最终语义特征记为 $SE(t)$, EEG 信号记为 x_i 。然后，通过以下损失函数优化语义编码器网络和全连接网络的权重：

$$Loss_S(\theta_S) = \lambda_1 L_C(SE(t), E_S(x_i)) + (1-\lambda_1) MSE(SE(t), E_S(x_i))$$

其中， x_i 表示 EEG 信号样本。

该损失函数由两部分组成：CLIP 损失 $L_C(\cdot,\cdot)$ 和均方误差（MSE）损失 $MSE(\cdot,\cdot)$ 。

CLIP 损失用于使一批图像标签的语义特征与相应视觉刺激下收集的 EEG 数据的编码结果之间的余弦距离尽可能小，并使不匹配特征的余弦距离尽可能大；而 MSE 损失用于使 EEG 数据的编码结果与语义特征之间的绝对距离尽可能接近。具体而言，我们设置超参数 $\lambda_1 = 0.8.$。采用 Adam 优化器训练网络，权重衰减设置为 0.0001，批次大小为 16，学习率为 0.001，训练轮次设置为 150。

2.2.2 视觉特征

预训练 VGG-19 模型的输出维度为 1000。通过将训练集中的所有图像输入该网络，我们获得了训练图像的视觉特征。在训练视觉编码器网络之前，我们进行主成分分析（PCA），提取前 20 个特征，这些特征贡献了原始特征 80.01%的方差。

视觉编码器模型的结构和损失函数与语义编码器相似。设图像为 i，PCA 后的图像潜在特征记为 $IE(i)$. 损失函数如下：

$$Loss_V(\theta_V) = \lambda_2 L_{CD}(IE(i), E_V(x_i)) \\ +(1-\lambda_2) MSE(IE(i), E_V(x_i)).$$

2.2.3 基于 EEG 的特征预测

训练两个编码器网络后，我们可以从 EEG 信号中预测语义特征和视觉特征。对于语义特征，将 EEG 信号输入语义编码器网络，得到相应的语义特征，用于两个目的：一方面，映射到 Glide 嵌入空间，作为扩散过程的文本引导信息，使扩散模型生成与语义输入匹配的图像；另一方面，通过训练一个线性映射层作为分类器，可以从语义特征中预测原始刺激图像的类别。同时，将 EEG 信号输入视觉编码器网络，得到相应的视觉特征，用于后续选择图像作为扩散过程的初始条件。

2.4 基于扩散模型的图像生成

为执行通用视觉解码任务，我们使用扩散模型生成图像，这比仅解码类别更进一步。基于从 EEG 信号中提取和预测的特征，采用预训练 Glide 模型进行图像生成，时间步长设为 100，引导尺度设为 7.0。

从 EEG 的语义特征中获得预测的图像类别，并将训练集中所有对应此类别的图像作为候选图像。EEG 的视觉编码器网络已通过对比学习进行优化，因此学习到的特征可用于衡量与目标特征的相似性。计算视觉特征与所有候选图像特征之间的向量余弦相似度，选择结果最大的图像作为初始图像。向初始图像添加 80 个时间步的噪声，作为扩散过程的初始噪声，随后执行 80 个时间步的反向扩散过程。

数据集中的图像均为物体照片。为使生成的图像更接近数据集中的图像，我们在提示词前添加前缀“a photo of”，输入扩散模型。然后，这些提示词将被转换到 Glide 嵌入空间，并生成文本条件图像。前缀“a photo of”在文本嵌入空间中对应一个维度为[3,512]的张量。数据集中所有标签对应的最大维度为[6,512]，因此我们通过添加填充标记将所有标签转换为[6,512]形状。随后，训练一个全连接网络，将 EEG 语义表征映射到标签嵌入中，再将其与前缀张量拼接，得到 Glide 文本嵌入。在反向过程中，将其作为条件输入 Glide 模型，执行无分类器引导。

3 实验

3.1 数据集与实现

EEG 数据集是在六名受试者观看视觉刺激图像时收集的。实验中使用的图像包括来自 ImageNet 数据集的 2000 张图像（40 类物体，每类 50 张图像）。图像按顺序展示，每张展示 0.5 秒；每类图像持续展示 25 秒，两类图像之间展示 10 秒的黑色图像。EEG 信号包含 128 个通道。参考以往研究，我们选择伽马波段的高频部分（55Hz 至 95Hz），该波段与视觉任务中的感知过程相关，且能实现最高性能。

在记录 EEG 信号的 500 毫秒内，我们剔除了前 40 毫秒和后 20 毫秒的部分，从剩余的 440 毫秒中截取连续的 160 毫秒记录，作为训练和解码过程的输入数据。所有通道的数据均用于语义提取。为更高效地提取视觉特征，我们选择了枕叶区域的 14 个通道（PO7、PO3、POz、PO4、PO8、POO9、POO1、POO2、POO10、O1、Oz、O2、Ol1h、Ol2h）。这些通道靠近 V1 视觉皮层，与低层视觉特征相关。

数据集按图像分为三个集合：80%用于训练，10%用于验证，10%用于测试，确保不同集合的图像在受试者之间不重叠。

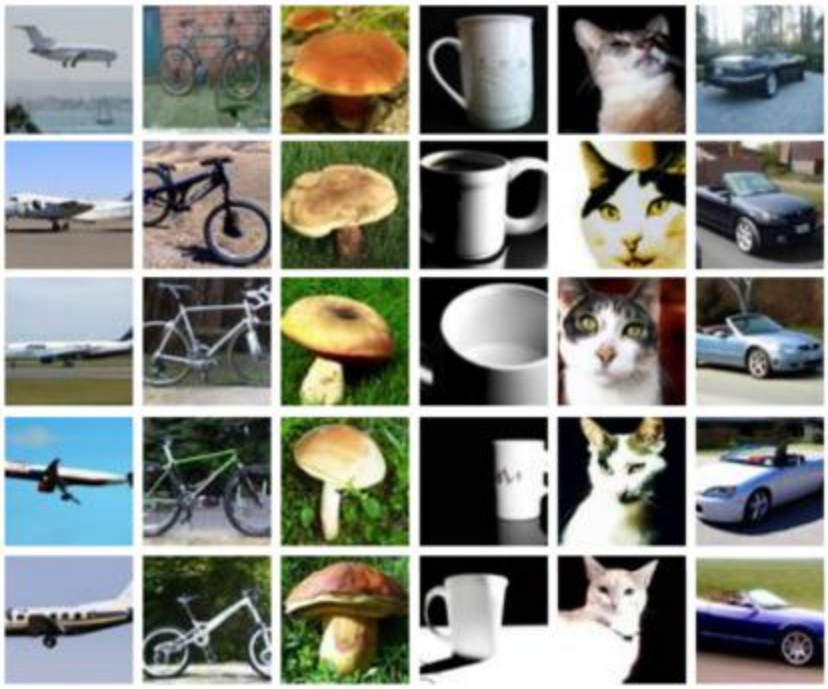
3.2 实验细节

本研究的架构包含 2 个编码器：视觉编码器有 1.9 万个参数，语义编码器有 52.3 万个参数；映射模型有 160 万个参数；扩散模型有 3.85 亿个参数，总时间步长为 160。推荐的硬件环境与我们训练和测试所使用的环境一致：CPU 为 Intel Xeon Gold 5115，GPU 为 NVIDIA GeForce RTX 3090Ti，内存为 32GB DDR4。

3.3 图像重建结果

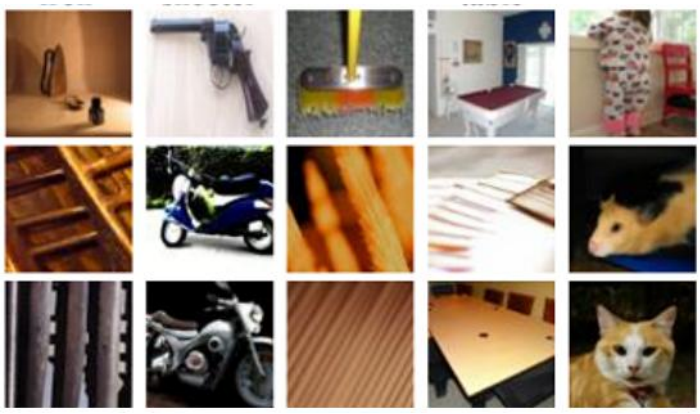
通过我们的框架，可以完成从 EEG 信号到图像的端到端重建。生成图像的部分示例如下图所示。可以看出，通过从 EEG 信号中提取语义特征和视觉特征，我们的方法重建的图

像在语义和视觉特征上均重现了刺激图像。对于同一真实图像，生成了不同的样本，表明我们的框架生成的图像具有高度多样性，而非简单重复相同结果。



第一行是真实标签，第二行是真实图像，其余行是不同的生成图像样本。

当然，也存在一些失败的结果。除了从 EEG 中提取错误特征外，有限的预训练生成模型也可能导致重建失败。下图展示了一些示例。由于我们使用预训练扩散模型作为生成模型，即使直接输入真实标签，它有时也无法生成正确的图像。这些失败结果表明了预训练扩散模型的不足，这可能会影响图像重建的准确性。



从上到下四行分别为：ImageNet 描述中的真实标签、视觉刺激的真实图像、直接输入真实标签生成的图像、输入从 EEG 信号中预测的特征生成的图像。

4．相关工作对比

4.4.2 图像重建

采用以下三个指标评估重建图像的质量：

- inception 分数（IS）：用于评估生成模型生成图像的质量和多样性，分数越高表明生成的物体越具体，不同类别的图像多样性越高。
- 结构相似性指数（SSIM）：用于量化两幅图像之间的相似性，通过考虑像素强度的局部模式及其关系来比较图像的结构信息。
- CLIP 分数（CS）：用于衡量图像与文本之间的语义相似性，也可通过计算两幅图像的 CLIP 嵌入的余弦相似性来评估图像之间的语义相关性。

我们计算了生成图像与真实图像之间的 CLIP 分数和 SSIM，以及每个类别的平均 inception 分数。表 3 展示了我们的方法与其他方法的定量结果对比。

表 3. 与其他方法的结果对比

Metrics	IS	CS	SSIM
Brain2Image[17]	5.07	/	/
NeuroVision[18]	5.15	/	/
Ours	**7.20**	**0.6230**	**0.1849**

为进一步展示我们的图像重建质量，我们从重建结果中选择了与其他方法展示结果相同类别的图像，可视化结果如下图所示。可以看出，我们的方法重建的图像在清晰度、真实感和多样性方面均更优。

从上到下三行分别为：Brain2Image、DreamDiffusion 和本文方法重建的图像。结果表明，我们的方法在低层和高层指标上均优于其他方法。

5. 结论

本方法提出了一种用于 EEG 表征学习和视觉解码的方法，高性能地完成了从 EEG 信号到图像的端到端重建任务。为缓解 EEG 视觉解码任务中面临的信号模式复杂、信噪比低等问题，探索了 LLMs 在这些任务中的潜在应用。我们将预训练 LLaMa-2 的隐藏状态作为

额外知识，以提升 EEG 表征学习和视觉解码的性能。通过对比学习，使 EEG 信号的表征与刺激图像在语义特征和视觉特征上对齐，高效地从脑活动中提取信息，最终重建出相应的刺激图像。为生成在低层和高层特征上均与真实图像相似的图像，我们在图像生成阶段通过预训练扩散模型融合语义特征和视觉特征，根据提取的特征从训练数据集中选择图像作为初始输入，并执行扩散过程。在 ImageNet-EEG 数据集上，我们的方法在 EEG 语义分类和图像重建方面均取得了当前最优结果。同时，本研究也存在一些局限性：例如，方法在其他数据集上的泛化能力尚不明确，对预训练模型的依赖性较强，且需要标注良好的数据集。这些问题需要在未来的 EEG 视觉解码任务中进一步探索。