基于 fMRI 信号的自然图像多标签分类研究

摘要

随着神经影像学与人工智能技术的迅速发展，脑-机接口和大脑信号解码成为了神经科学与人工智能交叉领域的重要研究方向。本文旨在探讨基于功能性磁共振成像（fMRI）信号在多标签图像分类任务中的表现。实验以被试观看自然图片时采集的 fMRI 信号为基础，构建多标签分类模型，用于预测图片中是否存在人物（person）、动物（animal）或飞机（airplane）三种语义类别。本研究提出并对比了两种不同的建模策略：一种为基于对比学习的 fMRI 特征增强模型，另一种为直接融合 fMRI 与图像特征的联合分类模型。研究表明，对比学习模型与融合模型各具优势，对比学习更善于从 fMRI 中提取结构清晰、表达稳定的语义特征；而融合模型则在多模态任务中具有更强的整体判别能力，在今后的工作中，可以考虑引入混合式策略。

1 引言 4
2 方法 4
2.1 数据结构与处理 4
2.2 模型设计 5
2.3 模型训练与评估 6
2.4 数据分析 6
3 结果 7
3.1 描述统计 7
3.2 对比学习模型 8
3.3 融合模型 15
3.4 两个模型的比较 21
4 讨论 23
4.1 模型结构与训练策略设计 23
4.2 学习率调整与收敛稳定性 23
4.3 模型性能差异与融合模型过拟合问题分析 24
4.4 特征空间与模型适用性 24
5 总结 25

1 引言

近年来，随着非侵入式神经影像技术（如 fMRI）的发展，研究人员得以以较高的空间分辨率观察人类大脑在感知、注意与认知任务下的活动模式。与此同时，深度学习在图像识别、语言处理等领域的广泛成功也激发了其在脑信号分析中的应用探索。将这两者结合，即通过深度学习模型解码 fMRI 信号中的潜在信息，有望实现更精确的脑功能识别甚至思维内容的解码。

在众多脑信号解码任务中，通过 fMRI 信号预测被试所观看的图片内容，是一项具有挑战性的研究课题。fMRI 数据具有高维、时序性强、噪声大等特点，如何从中提取有用的空间与时间特征，是建模过程中的关键问题。同时，在现实实验中，每幅图像往往包含多个语义类别，因而需要设计能够处理多标签输出的神经网络模型。

为此，本文提出并比较了两类建模方法：其一为对比学习方法，利用图像特征对 fMRI 特征空间进行监督，从而强化 fMRI 特征的语义对齐能力；其二为特征级融合方法，直接将 fMRI 特征与图像特征融合后进行分类预测。通过构建统一的数据加载、模型训练、评估与可视化流程，本文全面评估了这两类模型的优劣，并结合多种指标进行了性能分析。

2 方法

2.1 数据结构与处理

2.1.1 数据结构

实验使用的 fMRI 数据由被试在观看自然图片时采集，数据保存在 HDF5 文件中，分为训练集、验证集和测试集。每个数据样本包含以下内容：

voxel：经过预处理后的 fMRI 信号，数据类型为 float32，数据维度为(B, T, N)。其中 B 为样本数，T 为每张图片（每个试次）对应的时间点数量， N 为特征提取后的一维 fMRI 特征，此一维特征为每个体素做 GLM 分析得到的 beta

再进行标准化的值 ，这个一维的 beta 值是以图片类型为预测变量得到的 beta 值。

image：每个试次被试所观看的图片内容，数据类型为 float32，数据维度为(B, C, H, W)。其中 B 为样本数，C 为图片的 RGB 通道，H=224，W=224，H 和 W 为图片的高和宽（分辨率）

label：每张图片的类别标签，数据类型为 int64，数据维度为(B, 3)。其中 B 为样本数，3 表示有 3 个标签。该类别标签使用二进制分类（0/1 表示是否存在对应类别），标签的第 1 维代表了‘person’类别，第 2 维代表了‘animal’类别，第 3 维代表了‘airplane’类别。例如：[1, 1, 0]表示图片中存在‘person’和‘animal’， 但不存在‘airplane’。

2.1.2 数据处理

对 fMRI 数据进行了 NaN 检查与范围归一化处理。图像数据按照 ImageNet 均值与方差进行了标准化。对于训练数据中的类别不平衡问题，使用样本加权和 Focal Loss 进行缓解。在训练过程中设置了早停机制，当等待 25 个 epoch 无改变时，触发早停，并采用验证集 F1 分数作为模型选择的主要依据。初始设置的早停机制是以 loss 作为模型选择的主要依据，但是这样训练出来的模型在验证集上作检验的结果并不理想因此换成了 F1 分数。

2.2 模型设计

2.2.1 对比学习模型

该模型将图像作为监督信号，引导 fMRI 信号在特征空间中对齐图像语义。其结构包括，fMRI 编码器：输入为(T, N)格式的 fMRI 信号，采用 1D 卷积处理时间维度，线性层处理空间维度，再融合两者。图像编码器：以 ResNet18 为主干网络提取图像特征。对比学习模块：使用投影头对 fMRI 和图像特征进行映射，并计算 NT-Xent 对比损失。分类器：将 fMRI 特征输入多层全连接网络，输出三维的多标签预测。

该模型的训练目标由两个部分组成：主任务：多标签分类（使用 Focal

Loss）。辅助任务：fMRI 与图像对比学习（使用双向交叉熵损失）。

2.2.2 融合模型

融合模型同时提取 fMRI 和图像特征，并在特征级进行拼接，用于最终分类预测。fMRI 编码器：与对比模型相同。输入为(T, N)格式的 fMRI 信号，采用 1D 卷积处理时间维度，线性层处理空间维度。图像编码器：ResNet18+全连接层，输出与 fMRI 同维度的特征。融合机制：将两个模态的特征拼接（feature_dim×2），再输入分类器进行预测。该模型不包含对比学习损失，仅依赖分类损失进行优化。

2.3 模型训练与评估

训练策略：使用 AdamW 优化器与 CosineAnnealingWarmRestarts 学习率调度。每轮 epoch 记录训练/验证的 loss 与 accuracy。使用 macro F1 作为主要早停指标。Batch size 为 16。

评估指标：准确率（Accuracy）：预测正确的标签数与总标签数之比，衡量整体预测的正确性。精确率（Precision）：预测为正例的样本中，有多少是真正例，衡量模型输出的可信度。召回率（Recall）：所有真实为正例的样本中，被模型成功预测为正例的比例，衡量模型的漏报率。F1 分数（F1 Score）：精确率与召回率的调和平均，是分类效果的综合指标，尤其适合类别不平衡场景。AUC（Area Under Curve）：ROC 曲线下的面积，越接近 1 说明模型对正负样本区分越强。宏平均（Macro）：对每个类别分别计算指标后取平均，强调每类的重要性一致。微平均（Micro）：统计所有类别上的全局 TP、FP、FN 后计算指标，更关注整体样本分布。完全匹配率（Exact Match）：只有当所有标签都完全预测正确时才计为正确，反映模型多标签预测的完整性。

2.4 数据分析

本实验在 Python 3.11 环境下完成，主要使用的科学计算与深度学习库包括 PyTorch、NumPy、scikit-learn、Matplotlib 和 Seaborn 等。fMRI 信号与图像数据以 HDF5 格式存储，加载与预处理通过 h5py 和自定义 Dataset 实现。图像

部分在输入前统一 resize 为 224×224，并进行标准化处理。fMRI 数据维度为 (B, T, N)，在建模前对每个时间序列进行了归一化与时间点平均，最终提取出每个样本的一维 beta 表征用于输入模型。所有模型训练均在 GPU 环境下进行，支持 Early Stopping、动态学习率调整与性能监控，训练日志及中间结果均通过可视化工具进行记录与分析。

3 结果

3.1 描述统计

3.1.1. 数据检查

训练集：fMRI 数据形状: (8700, 3, 15724)，图像数据形状: (8700, 3, 224, 224)，标签数据形状: (8700, 3)，数据中均不包含 NaN。

验证集：fMRI 数据形状: (300, 3, 15724)，图像数据形状: (300, 3, 224, 224)，标签数据形状: (300, 3)，数据中均不包含 NaN。

测试集：fMRI 数据形状: (1000, 3, 15724)，图像数据形状: (1000, 3, 224, 224)，数据中均不包含 NaN。

3.1.2 标签统计

训练集：person: 正样本 4489, 负样本 4211 (51.6%正例)，权重设置: 0.94；animal: 正样本 1863, 负样本 6837 (21.4%正例)，权重设置: 3.67；airplane: 正样本 236, 负样本 8464 (2.7%正例)，权重设置: 35.86。

验证集：person: 正样本 147, 负样本 153 (49.0%正例)，权重设置: 1.04；animal: 正样本 71, 负样本 229 (23.7%正例)，权重设置: 3.23；airplane: 正样本 8, 负样本 292 (2.7%正例)，权重设置: 36.50。

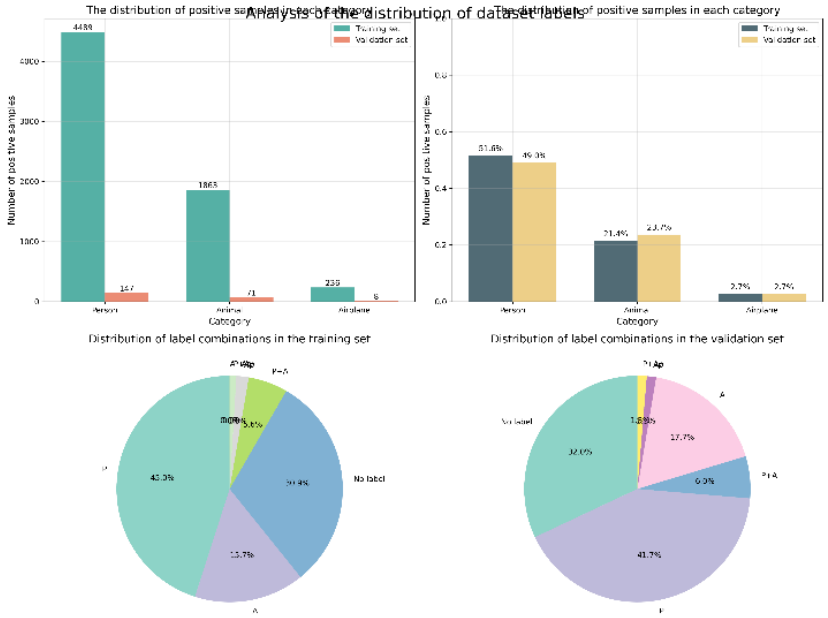

图 1 训练集和验证集标签分布

如图 1 所示，三类标签在两个数据集中均存在显著的不均衡现象。下方的饼图进一步给出了多标签组合的分布。在训练集中，单一标签‘person’占比为 45.0%，‘animal’为 15.7%；‘person+animal’组合占比为 5.6%；无标签样本占 30.9%；其余组合（如‘person+airplane’、‘全部标签’等）比例较低。验证集的分布与训练集相似，‘person’标签样本占比 41.7%，‘animal’标签为 17.7%，‘person+animal’组合占 6.0%，无标签样本为 32.0%。

3.2 对比学习模型

对比学习模型先采用 Resnet18 来提取图像的特征，接着编写了一个 CNN 模型来对 fMRI 的 1D 时间序列进行卷积处理，使用投影头对 fMRI 和图像特征进行映射，最后将 fMRI 特征输入一个全连接层，输出标签预测。在训练集进行训练的时候使用了’voxel’, ‘image’和‘label’， ‘image’在训练集中起到一个监督信号的作用，将 fMRI 信号与图像特征对齐，在验证集上使用‘voxel’预测‘label’并与真实‘label’比较检验对比学习模型的性能，测试集也是通过训练好的模型使用‘voxel’来预测生成‘label’。进行模型在验证集上的性能评估以及将模型应用到测试集上生成预测标签，结果如下：

3.2.1 训练过程

数据加载器的配置为：批次大小: 16；训练批次: 544 (使用平衡采样)；验证

批次: 19；测试批次: 63。总 epoch 数为 100，初始学习率为 0.001，训练了 45 轮后触发了早停，选取了 epoch=20 时的模型，该模型的最佳验证 F1=0.9034，此时的学习率被调整为 0.001。

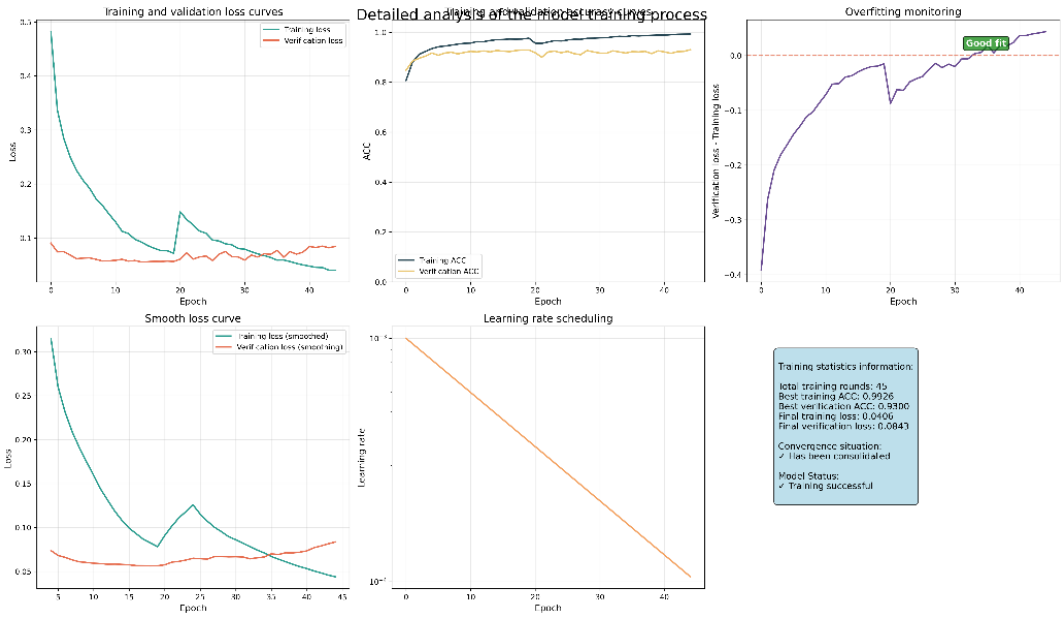

图 2 训练过程详细信息

根据上图显示，在使用对比学习模型训练了 45 轮后，触发早停，选择训练 20 轮时的模型为最佳模型。经检验模型收敛，同时添加过拟合监控也没有发现模型出现过拟合的情况。选择的最佳模型的训练集的准确率为 0.9926，验证集的准确率为 0.93，最终的训练集的损失为 0.0406，验证集的损失为 0.0843。

3.2.2 验证集性能评估

表 1 整体性能指标

准确率	宏平均 F1 分数	微平均 F1 分数	宏平均精确率	宏平均召回率	完全匹配率
0.9300	0.8453	0.8511	0.9439	0.7683	0.8100

对比学习模型在验证集上的整体表现如表 1 所示，准确率为 0.9300 表示模型在所有标签上的整体预测准确性已达 93%，说明大部分标签判断是正确的。宏平均 F1 分数为 0.8453 和微平均 F1 分数为 0.8511 均较高，显示模型不仅在样本整体层面有良好表现（微平均），在每个类别的平衡表现上也有保障（宏平

均）。宏平均精确率为 0.9439 远高于宏平均召回率（0.7683），说明模型更擅长正确识别正例的同时，可能遗漏了一部分应识别的正例，即有一定的漏报现象。完全匹配率为 0.8100 表示有 81%的样本在三个标签上全部预测正确，考虑到是多标签任务，这一比例非常可观，说明模型具备较强的综合识别能力。准确率和完全匹配率的区别是准确率是允许标签预测值与真实值不完全对应，允许存在部分准确，完全匹配是指需要标签中三个类别都完全对应才判为正确，不允许部分准确存在，是更为严格的判定指标。

表 2 各类别详细指标

类别	准确率	精确率	召回率	F1	AUC
Person	0.8767	0.9044	0.8367	0.8693	0.9403
Animal	0.9200	0.9273	0.7183	0.8095	0.9384
Airplane	0.9933	1.0000	0.7500	0.8571	0.9649

对于每个类别的具体预测表现如表 2 所示：对于 Person 类别：准确率（0.8767）与 精确率（0.9044）均表现良好，说明模型对于该类别的判断非常可靠。召回率（0.8367） 略低，意味着仍有一部分“person”实例未被检测出来。F1 分数（0.8693）综合了精确率与召回率，是三类中表现最稳的。AUC（0.9403）显示模型在该类别上的分类能力非常强；对于 Animal 类别：准确率（0.9200）和精确率（0.9273）均很高，但召回率（0.7183）较低，说明模型对‘animal’标签存在一定的漏检现象。F1 分数（0.8095）仍处于较高水平，但已明显低于‘person’类别。AUC（0.9384）同样表明模型有很强的判别力，但需要改进召回策略；对于 Airplane 类别：准确率（0.9933） 与 精确率（1.0000）极高，说明模型在预测‘airplane’为正例时几乎没有误报。召回率（0.7500）表示只有 75%的正例被成功识别，结合极高的精确率，说明模型偏向于“保守”预测，即宁可漏检也尽量不误判。F1 分数（0.8571）在小样本类别中表现优异。AUC（0.9649）是三类中最高的，验证了即使数据极不平衡，模型在 airplane 类别上依然有较强的判别能力。

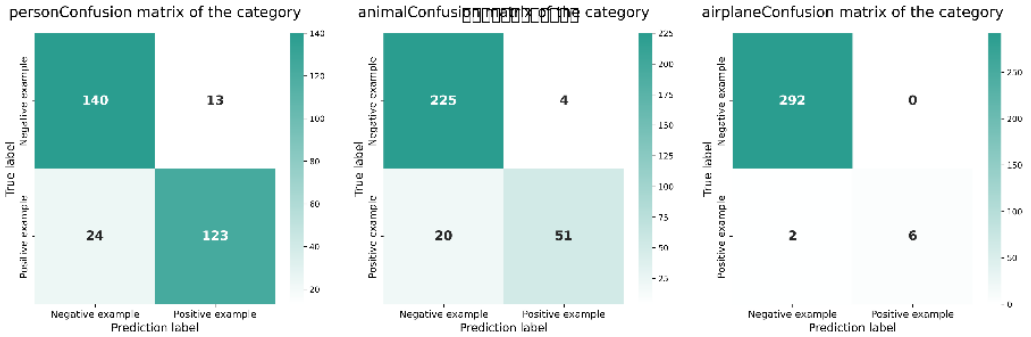

图 3 各类别混淆矩阵

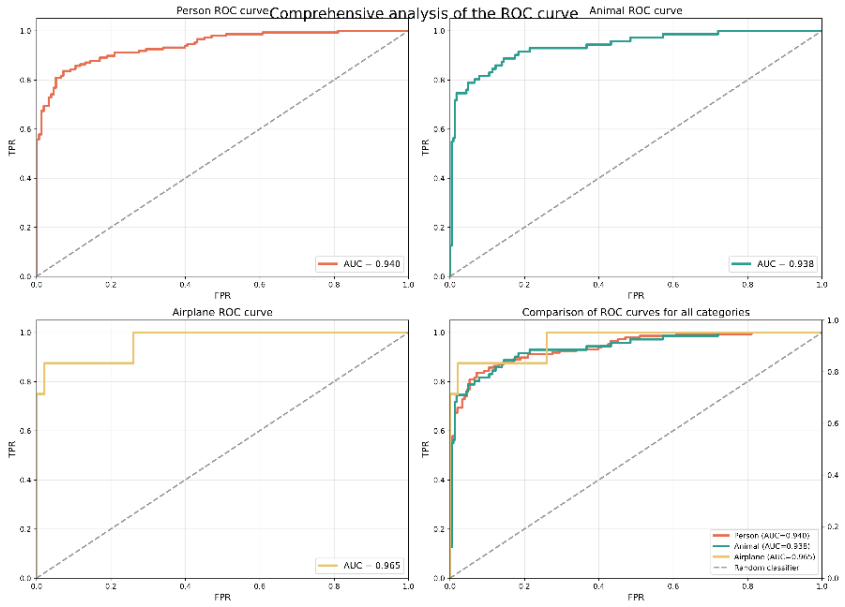

图 4 ROC 曲线

图 3、4 展示的是三个类别的混淆矩阵及其对应的 ROC 曲线。对于 Person 类别，真阳性（TP）= 123，假阴性（FN）= 24，假阳性（FP）= 13，真阴性（TN）= 140 模型对‘person’类别表现出较高的精准率与良好的召回，误判和漏判数量均处于可控范围，整体预测表现稳定，同时 ROC 曲线呈现良好的左上弯曲走势，AUC 值达到 0.940，说明模型在该类别的真阳性率高，同时假阳性率控制良好，具备较强的二分类判别能力；对于 Animal 类别 TP = 51，FN = 20，FP = 4，TN = 225，假阳性数量极低，仅为 4 例，显示出极高的精确率，但假阴性仍占据较大比例，表明模型偏向保守判断，召回率受到影响，AUC 为 0.938，曲线走势与‘person’类别接近，表明模型对动物类别的识别能力同样优秀，尽管在后续混淆矩阵中我们会看到召回率略低，但总体模型区分能力强；对于 Airplane 类别 TP = 6，FN = 2，FP = 0，TN = 292，模型对该类别几乎没有

误报（FP = 0），实现了 100%精确率，但漏检两例仍存在，考虑到类别本身样本极少，该表现仍属优秀，尽管该类别样本数量稀少，模型依然获得了最高的 AUC 值（0.965），这表明其在极端类别不平衡条件下，依旧能够捕捉到有效的区分边界。不过应注意的是，AUC 较高并不等同于整体 F1 分数最高，因为该指标不考虑分类阈值固定后的精度与召回平衡。

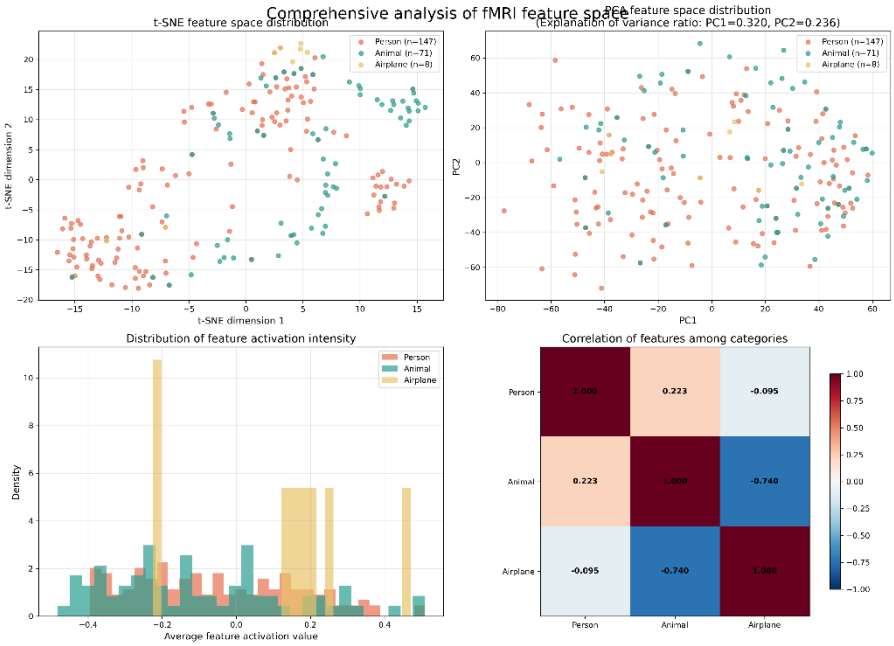

图 5 验证集 fMRI 特征空间分布

上图展示了模型在验证集上提取的 fMRI 特征空间的综合可视化结果，包括 t-SNE 分布、PCA 分布、激活强度统计及类别间相关性分析。从左上角的 t-SNE 图可见，模型提取的高维 fMRI 特征在低维空间中形成了较清晰的类别分布，其中‘person’类别聚类最明显，‘animal’类也呈现部分集中，而‘airplane’由于样本稀少，分布较为稀疏但与其他类别边界清晰。右上角的 PCA 图表明，在解释方差最高的两个主成分上，虽然类别间存在较多重叠，但‘airplane’类依然展现出相对独立的分布趋势。左下角的特征激活强度直方图显示，‘airplane’类别的特征激活分布最集中，说明模型对其提取的表示较为一致；而‘person’与‘animal’的激活范围更广，反映出更复杂的响应模式。右下角的类别间相关性热力图进一步揭示，‘person’与‘animal’之间具有一定程度的正相关（0.223），说明二者在大脑激活模式中存在一定共享区域，而‘airplane’与两者均呈负相关（-0.095 和 -0.740），说明该类在 fMRI 表征空间中具有明显的区分性。整体来

看，该图反映出模型能够在 fMRI 特征层面有效捕捉不同语义类别的神经响应差异，具备良好的表示能力和类内一致性。

3.2.3 测试集预测标签

表 3 标签结果统计

类别	预测为正例数量（n=1000）	预测为正例概率（%）	平均概率
person	413	41.3	0.439
animal	229	22.9	0.254
airplane	17	1.7	0.033

表 4 标签组合分布

组合	样本数（n=1000）	概率（%）
person	381	38.1
无标签	374	37.4
animal	199	19.9
person+animal	29	2.9
airplane	14	1.4
person+airplane	2	0.2
person+animal+airplane	1	0.1

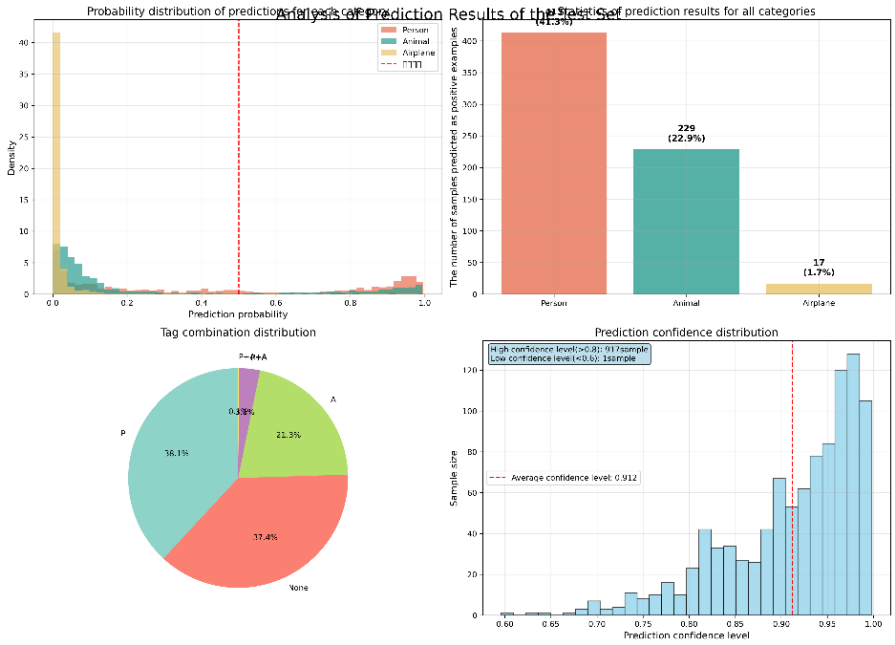

图 6 测试集预测标签的统计分布

上面的表 3、4 和图 6 展示了模型在测试集上的预测结果统计、预测概率分布、标签组合情况及置信度分布。

从图 6 右上角柱状图及配套表格可见，在 1000 个测试样本中，模型分别将 41.3%的样本预测为包含‘person’类，22.9%预测为‘animal’类，1.7%预测为‘airplane’类。这一分布与训练/验证集中真实标签分布基本一致，表明模型在预测阶段保持了类别敏感性的延续。此外，三类标签对应的平均预测概率分别为 0.439（person）、0.254（animal）和 0.033（airplane），说明模型在对不同类别进行打分时具备显著的置信度差异。

图 6 左上角的预测概率分布图进一步揭示了这一趋势——大部分‘airplane’预测概率集中在低于 0.1 的区间，而‘person’与‘animal’类则在两端分布，具备更高的区分性。中间虚线标示了默认阈值 0.5 的划分位置，可见通过此阈值大致可将样本有效划分为正负类。

图 6 左下角的标签组合饼图显示，在测试集中，单一标签‘person’的预测组合最多，占 38.1%；其次是无标签样本，占比 37.4%；‘animal’标签单独出现的比例为 19.9%；其余多标签组合如‘person+animal’（2.9%）、‘airplane’（1.4%）以及三标签共现（0.1%）出现频率极低。整体预测结果仍以单标签与无标签为主，符合训练阶段标签分布。

图 6 右下角直方图展示了所有样本的预测置信度分布。模型对大多数样本的输出置信度集中在 0.8 至 1.0 之间，整体偏向高置信度输出。图中统计结果表

明，有 917 个样本的预测置信度高于 0.8，仅有 1 个样本低于 0.6，平均置信度为 0.912。这说明模型在测试集上的决策非常明确，大部分预测结果具有较高的可信度。

3.3 融合模型

融合模型先使用 Resnet18 提取图像特征，输出与 fMRI 信号同维度的特征，接着使用一个 CNN 模型来对 fMRI 的 1D 时间序列进行卷积处理，将两个模态的特征拼接后输入分类器进行标签预测。在训练集进行训练的时候使用了‘voxel’, ‘image’和‘label’，在验证集使用‘voxel’, ‘image’预测‘label’并与真实的‘label’进行比较检验模型性能，在测试集中无‘label’仅使用‘voxel’, ‘image’生成预测的‘label’。数据加载器的配置为：批次大小: 16；训练批次: 544 (使用平衡采样)；验证批次: 19；测试批次: 63。总 epoch 数为 100，初始学习率为 0.001，训练了 40 轮后触发了早停，选取了 epoch=15 时的模型，该模型的最佳验证 F1=0.8570，此时的学习率为 0.000147。进行模型在验证集上的性能评估以及将模型应用到测试集上生成预测标签，结果如下：

3.3.1 训练过程

数据加载器的配置为：批次大小: 16；训练批次: 544 (使用平衡采样)；验证批次: 19；测试批次: 63。总 epoch 数为 100，初始学习率为 0.001，训练了 40 轮后触发了早停，选取了 epoch=15 时的模型，该模型的最佳验证 F1=0.8570，此时的学习率被调整为 0.000147。

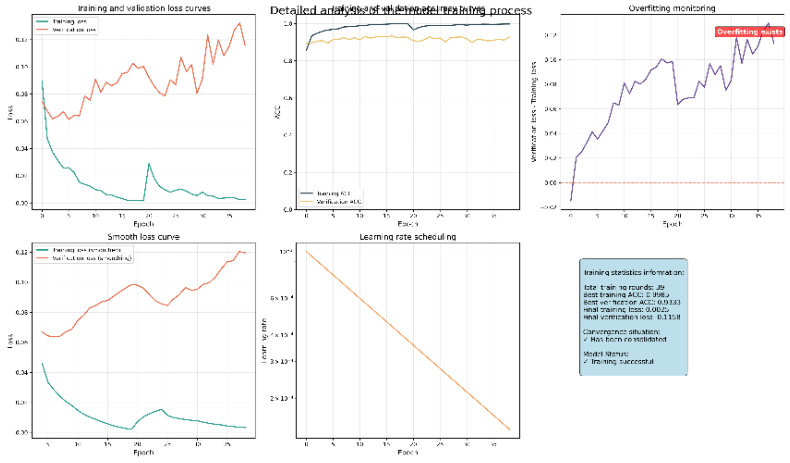

图 7 训练过程详细信息图

根据上图显示，在使用融合模型训练了 40 轮后，触发早停，选择训练 15 轮时的模型为最佳模型。添加过拟合监控发现模型出现过拟合的情况，但是经检验模型收敛。选择的模型的训练集的准确率为 0.9985，验证集的准确率为 0.9333，最终的训练集的损失为 0.0025，验证集的损失为 0.1158。

3.3.1 在验证集上进行性能评估：

表 5 整体性能指标

准确率	宏平均 F1 分数	微平均 F1 分数	宏平均精确率	宏平均召回率	完全匹配率
0.9289	0.7898	0.8505	0.9360	0.7038	0.8100

融合模型在验证集上的整体表现如表 5 所示，准确率为 0.9289，表示模型在所有标签上的整体预测准确性已达到 92.89%，说明多数标签判断是正确的。宏平均 F1 分数为 0.7898，微平均 F1 分数为 0.8505，整体而言模型在样本整体层面上（微平均）具有良好表现，而在各类之间的均衡性（宏平均）略低于对比学习模型。宏平均精确率为 0.9360，高于宏平均召回率（0.7038），表明模型倾向于输出较为保守的判断，即更擅长避免误报，但存在一定的漏检。完全匹配率为 0.8100，与对比模型一致，说明该模型有 81%的样本在三个标签上完全预测正确，体现了良好的多标签整体识别能力。准确率与完全匹配率的差异在于前者允许部分标签正确，而后者要求三个标签全部预测正确，是更严格的判定标准。

表 6 各类别详细指标

类别	准确率	精确率	召回率	F1	AUC
Person	0.8700	0.8913	0.8367	0.8632	0.9285
Animal	0.9300	0.9167	0.7746	0.8397	0.9507
Airplane	0.9867	1.0000	0.5000	0.6667	0.9756

对于每个类别的具体预测表现如表 2 所示：对于 Person 类别，准确率为 0.8700，精确率为 0.8913，表明模型对该类有较高识别稳定性；召回率为

0.8367，说明模型在一定程度上仍有漏检现象。F1 分数为 0.8632，为三类中最高值，显示在该类别上表现最为均衡。AUC 为 0.9285，说明模型在该类别具备较强的区分能力；对于 Animal 类别，准确率为 0.9300，精确率为 0.9167，显示模型在避免误报方面表现出色；召回率为 0.7746，表明仍有部分‘animal’样本未被识别。F1 分数为 0.8397，略低于‘person’类，但仍处于较高水平；AUC 为 0.9507，为三类中第二高，表明该类别下模型具有良好的分类能力；对于 Airplane 类别，准确率为 0.9867，精确率为 1.0000，说明模型在预测为正例时几乎从不误报；但召回率仅为 0.5000，表示该模型在一半的“airplane”样本上未能识别出目标，呈现出更强的保守性。F1 分数为 0.6667，为三类中最低，受限于召回率不足；但 AUC 仍达 0.9756，为所有类别中最高，说明模型具备极强的判别边界，但实际预测中采用了较高的判断阈值。

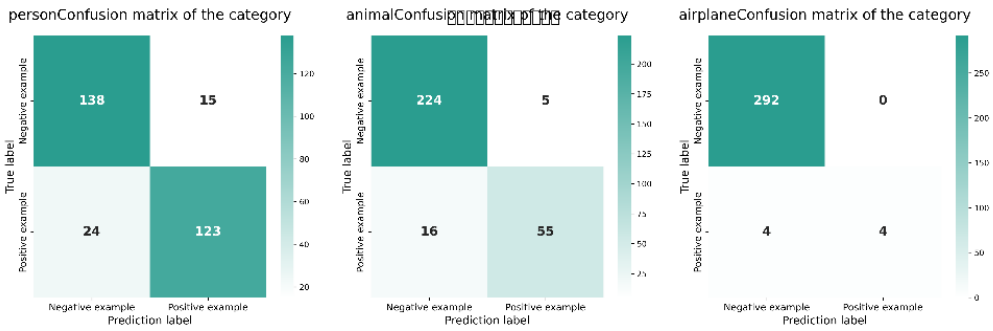

图 8 各类别混淆矩阵

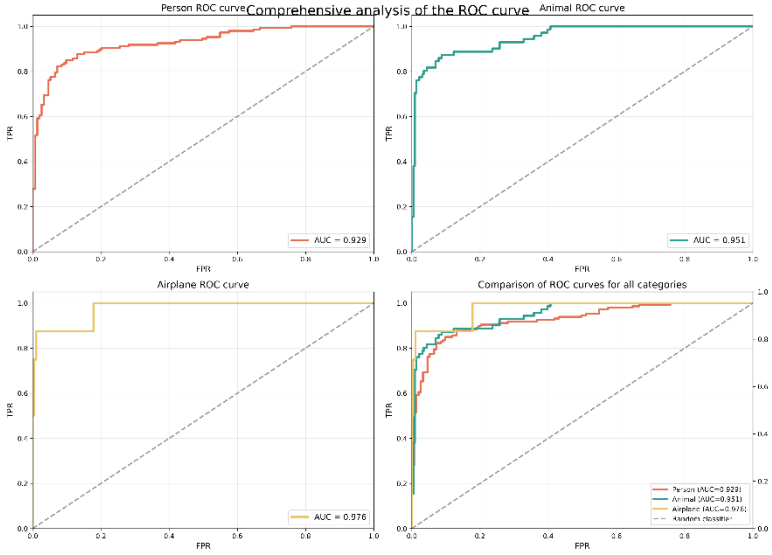

图 9 ROC 曲线

图 8、9 展示了融合模型在三个类别上的混淆矩阵及其对应的 ROC 曲线。对于 Person 类别，真阳性（TP）= 123，假阴性（FN）= 24，假阳性（FP）= 15，真阴性（TN）= 138，模型在该类别上的召回率和精确率表现均衡，尽管存在少量误报与漏报，整体识别性能依然稳定。ROC 曲线呈现理想的左上走势，对应 AUC 值为 0.929，表明模型在该类别上的判别能力较强，能够有效区分正负样本。对于 Animal 类别，TP = 55，FN = 16，FP = 5，TN = 224，假阳性数量控制良好，仅为 5 例，显示模型在识别负样本方面具有较强能力。假阴性数量略高，说明在部分‘animal’样本的识别上仍有漏检现象。该类别对应的 AUC 值为 0.951，显示模型具备极高的区分能力，ROC 曲线与‘person’类别相似，说明整体判断较为可靠。对于 Airplane 类别，TP = 4，FN = 4，FP = 0，TN = 292，模型在该类别上同样实现了 100%精确率（无误报），但召回率仅为 50%，即在全部 8 个正例中仅识别出 4 个。这说明模型在面对极小样本类别时采取了更为保守的策略，倾向于不做正例判断以避免误报。尽管如此，该类别的 ROC 曲线 AUC 达到了 0.976，为三个类别中最高，表明其在不同阈值下的区分边界最为清晰，进一步说明模型具备出色的判别潜力，尤其是在面对不平衡类别时。整体来看，融合模型在三个类别上均展现出较高的 AUC 值和稳定的预测性能，尤其在高精确率与低假阳性控制方面表现良好，验证了其在 fMRI 与图像融合特征下的判别能力。

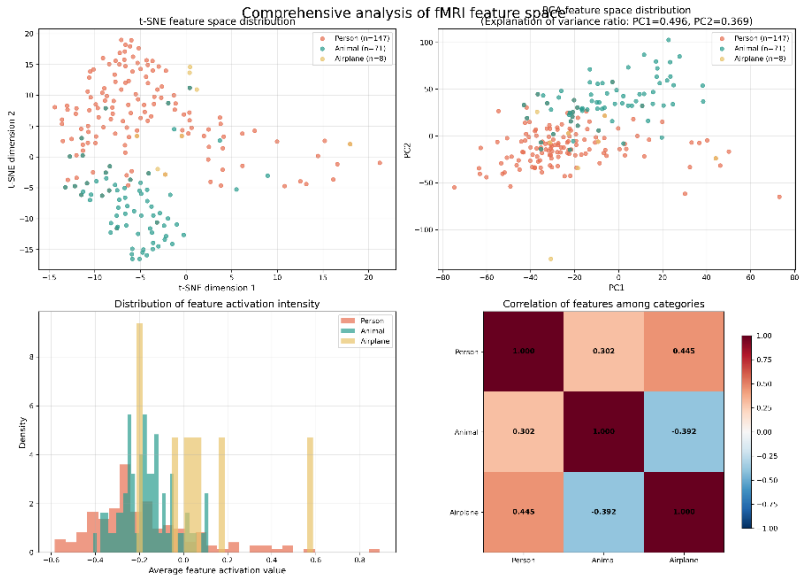

图 10 验证集 fMRI 特征空间分布

高维 fMRI 特征在降维后形成了具有一定分离性的类别分布，其中‘person’类别呈现较明显的聚类结构，‘animal’类别也形成了相对集中分布，而‘airplane’类由于样本极少，分布较为离散，但仍呈现一定边界清晰度。右上角的 PCA 分布图表明，在前两个主成分维度（PC1=49.6%、PC2=36.9%）构成的空间中，不同类别存在较大程度的重叠，但‘airplane’样本仍保持一定的独立性，说明该类特征在主方向上保留了区别性信息。左下角的特征激活强度直方图展示了三类样本的平均特征值分布情况，其中‘airplane’类别激活分布较集中，表明模型对该类别的表征更为一致；而‘person’与‘animal’类别的激活值分布更为广泛，反映出神经响应模式的多样性。右下角的类别间相关性热力图显示，‘person’与‘animal’类别之间存在中等程度正相关（0.302），而‘airplane’分别与‘animal’呈负相关（-0.392），与‘person’呈中等正相关（0.445），说明其在表征空间中同时包含一定相似性与差异性。整体来看，该图表明模型能够在特征层面较好地刻画三类语义刺激的神经响应模式，具备较强的类别分离能力与表达一致性。

3.3.3 测试集预测标签

表 7 标签结果统计

类别	预测为正例数量（n=1000）	预测为正例概率（%）	平均概率
person	421	42.1	0.423
animal	209	20.9	0.221
airplane	15	1.5	0.021

表 8 标签组合分布

组合	样本数	概率（%）
person	403	40.3
无标签	373	37.3
animal	193	19.3

person+animal	16	1.6
airplane	13	1.3
person+airplane	2	0.2

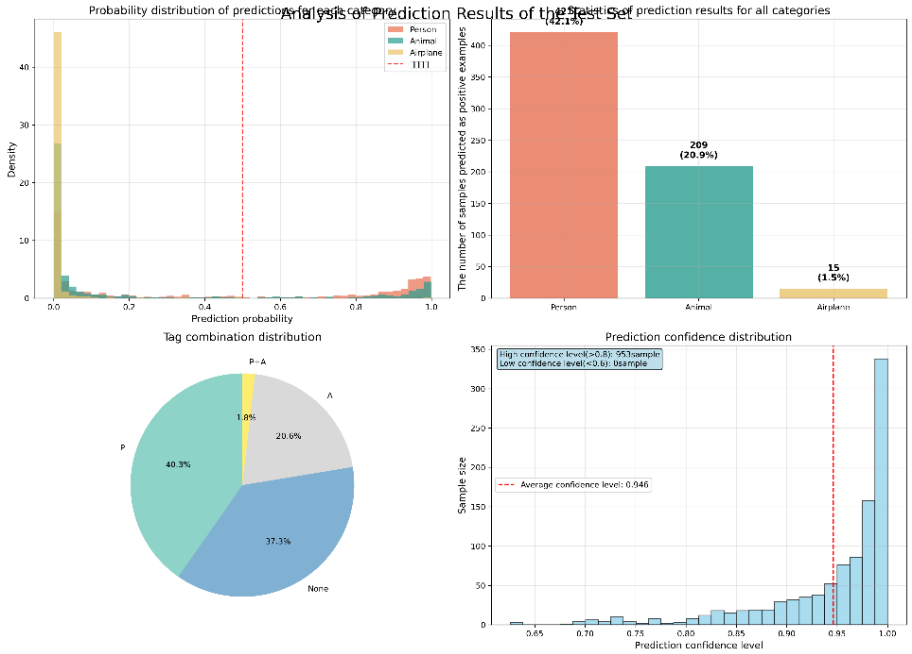

图 11 测试集预测标签的统计分布

上面的表 7、表 8 和图 11 展示了模型在测试集上的预测结果统计、预测概率分布、标签组合情况及置信度分布。

从图 11 右上角柱状图及配套表格可见，在 1000 个测试样本中，模型分别将 42.1%的样本预测为包含‘person’类，20.9%预测为‘animal’类，1.5%预测为‘airplane’类。整体分布与训练/验证集中的类别比例基本一致，说明模型在推理阶段仍保持了较强的类别敏感性。三类标签对应的平均预测概率分别为 0.423（person）、0.221（animal）和 0.021（airplane），表明模型在打分过程中对不同类别保持了明显的置信度区分，尤其对“airplane”类呈现出极低的预测激活水平。

图 11 左上角的预测概率分布图进一步揭示了这一趋势：大多数‘airplane’类别的预测概率集中在极低区间（小于 0.1），而‘person’与‘animal’类的概率分布相对更广，分布在接近 0 或 1 的两端，表明模型能够较好地区分正负样本。图

中红色虚线标示默认阈值 0.5，基本能够有效划分样本预测类别。

图 11 左下角为标签组合饼图，显示在测试集中，‘person’标签的单独预测组合最多，占比 40.3%；其次是无标签样本，占比 37.3%；‘animal’类单标签样本占 19.3%；其余组合如‘person+animal’（1.6%）、‘airplane’（1.3%）和‘person+airplane’（0.2%）占比较低。整体预测结果仍以单标签和无标签为主，符合训练数据的标签结构。

图 11 右下角直方图展示了所有样本的预测置信度分布情况。模型对大部分样本的输出置信度集中在 0.9 以上，显示出较强的判断明确性。统计结果显示，有 953 个样本预测置信度高于 0.8，无样本低于 0.6，整体平均置信度达到 0.946，说明模型在测试集上做出的预测普遍可信，置信水平较高。

3.4 两个模型的比较

本研究设计并比较了两类模型：一是基于 fMRI 与图像特征对齐的对比学习模型，二是在分类阶段融合 fMRI 和图像双模态输入的融合模型。以下从数据使用策略、分类性能、特征表达、类别不平衡鲁棒性以及训练过程稳定性五个方面对两种模型的优势与不足进行分析。

3.4.1 数据利用与研究目标

从研究目标来看，本研究旨在通过 fMRI 信号预测被试观看图像的类别标签。对比学习模型训练阶段引入图像特征进行对齐优化，在测试阶段仅依赖 fMRI 信号进行预测，未直接利用图像信息。这种设计虽然体现了更强的神经信息解码能力，但未充分利用原始图像数据的丰富视觉线索。融合模型在训练与测试阶段均同时接入 fMRI 和图像特征，增强了对多模态数据的整体利用率。

3.4.2 性能对比

从验证集整体评估指标来看，对比学习模型在准确率（0.9300）与 F1 分数（宏平均 0.8453，微平均 0.8511）方面略高于融合模型（准确率 0.9289，宏平均 F1 为 0.7898），显示出在整体分类能力上更优。

在各类别细节指标中：对比学习模型在‘Person’类别上的 F1 分数为 0.8693，融合模型为 0.8632，性能相近；在‘Animal’类别上，对比模型 F1 为 0.8095，融合模型提升至 0.8397，融合策略在该类更具优势；对于少数类‘Airplane’，对比模型 F1 达到 0.8571，明显高于融合模型的 0.6667，显示其在小样本条件下表现更稳定。从 AUC 指标看，两模型在各类别均表现良好，对比模型在‘person’与‘airplane’类别中略胜一筹，融合模型则在‘animal’与‘airplane’上获得更高的 AUC（0.951 和 0.976），说明其在不同阈值下的判别能力强。

3.4.3 特征对齐与多模态融合优势互补

从特征空间可视化（t-SNE 与 PCA）和特征激活图来看：对比学习模型对 fMRI 表征的分离性更强，尤其在‘person’和‘airplane’类别上体现出较好的类间边界，说明其在对神经响应特征的细粒度区分与对齐上具有优势；融合模型的特征分布更为紧凑‘animal’ 与‘person’类别间的相关性更高（如相关性热力图中 Person-Airplane 达 0.445），表明其在整合图像特征后形成了更一致的跨模态表达结构。因此，两者在模型结构和特征建模方面各有侧重：对比学习在 fMRI 表征提取与对齐精度上更优，而融合模型则有利于整合多模态信息以增强总体判别能力。

3.4.4 训练过程的稳定性与收敛性分析

从训练损失与准确率曲线来看：对比学习模型的训练损失下降速度较快，准确率曲线在前期收敛迅速，整体波动较小，显示其在对比学习目标指导下能快速稳定地构建有效表征；融合模型训练过程略显不稳定，尤其在多模态特征融合初期，模型对不同信息源的权重学习存在一定波动，整体准确率提升相对缓慢。这说明对比学习模型在优化过程中具备更好的收敛特性，而融合模型在训练早期可能需要更精细的正则策略或对抗训练手段以控制训练噪声。

3.4.5 类别不平衡条件下的表现

在面对严重类别不均衡的测试集中，两个模型在‘airplane’类别上均呈现出较强的区分能力（AUC 均超过 0.96），但表现机制不同：对比模型在‘airplane’类的召回率较高（0.75），F1 达到 0.8571，说明其在小样本类别识别上具有更好

的覆盖能力；而融合模型虽然具备极高的精确率（1.000）和更高的 AUC（0.976），但召回率仅为 0.500，说明其预测更保守，更易出现漏报。因此，在类别不平衡情境下，对比模型通过优化特征对齐策略在小类识别上保持了较高的稳健性，而融合模型可能受限于视觉主导特征而在弱类样本中表现保守。

4 讨论

本研究围绕基于 fMRI 信号预测图像语义标签的任务，设计并实现了对比学习模型与融合模型两种策略，并对其性能进行系统对比与分析。在模型开发与实验过程中，针对任务特点与数据限制，引入了多项设计与调控机制以增强模型的泛化能力与学习稳定性。

4.1 模型结构与训练策略设计

为更有效地建模大脑活动与视觉信息之间的对应关系，对比学习模型采用了“fMRI–图像特征”双塔结构进行特征对齐。在结构设计上，本研究选用了经典的 ResNet-18 作为图像编码器，其优势在于：网络深度适中、参数量适中、训练速度快，同时具备较强的图像语义提取能力，特别适合在数据规模较小或有限计算资源下的语义建模任务。

fMRI 编码器部分采用由线性层与激活函数构成的简单全连接网络，以防止在样本规模不大的情况下模型过拟合，同时通过标准化与正则化控制输出空间分布。实验中，两个模型均加入了 Dropout 和 L2 正则项，以减少参数冗余和缓解过拟合。

此外，在训练过程中，所有模型均引入了 Early Stopping（早停机制），即当验证集性能在连续若干 epoch 无明显提升时自动停止训练。我们观察到融合模型在训练初期很快达到较低训练误差，但验证性能随后开始波动甚至下降，提示其出现了过拟合倾向。早停机制在此处有效避免了模型在验证集性能上的退化，是应对小样本数据常用的防过拟策略之一。

4.2 学习率调整与收敛稳定性

学习率是神经网络训练中最关键的超参数之一。本实验使用了基于 epoch 的学习率衰减策略，初始学习率设置较高以便模型快速探索最优解空间，并在训练中后期逐步减小，以稳定优化过程并防止跳出最优点附近。实践证明，该策略对对比学习模型尤其有效，其损失下降平稳、收敛速度快，训练过程波动较小。

相比之下，融合模型在早期阶段存在较大梯度波动，可能由于 fMRI 和图像模态的表征方式差异较大，特征融合初期模型难以对两类输入建立平衡权重。这也提示后续工作中可进一步引入模态注意力机制、自适应加权或门控机制，以更有效地调控跨模态特征流动。

4.3 模型性能差异与融合模型过拟合问题分析

融合模型在整体分类准确率与 AUC 指标上表现良好，但在部分类别（尤其是 airplane）上召回率明显偏低，F1 分数也低于对比模型，体现出模型对小样本类别的过拟合与偏保守预测倾向。结合激活分布图和 ROC 曲线分析，我们推测这与图像模态对主类（如 person）的强引导作用有关，模型更倾向于依赖视觉显著类别特征进行判断，导致对罕见类缺乏泛化。此外，融合模型在训练集上 loss 下降迅速，验证集却早早触发 early stopping，进一步说明该模型在高维特征融合空间中存在过拟合风险。未来可以考虑使用数据增强、标签平滑、或重采样策略等方式来提升其对低频标签的鲁棒性。

4.4 特征空间与模型适用性

通过可视化特征空间（t-SNE、PCA）和类别相关性矩阵的分析，我们观察到对比学习模型更容易形成语义明确的类别聚类，尤其在 fMRI 特征上能够有效建立类内一致性与类间区分性。这体现了对比学习方法在表征层面上具备更好的结构约束能力，尤其适合于面向“表示学习”目标的神经数据建模任务。

相比之下，融合模型通过直接引入图像编码增强了整体判别能力，但在高层表达上反而可能降低了对 fMRI 信号内部结构的细粒度捕捉能力。这表明：对比学习方法在特征结构建模、类间边界优化方面仍具一定优势，而融合模型则更适合偏任务驱动、目标导向的应用情境。

5 总结

整体而言，本研究中的对比学习模型与融合模型各具优势，对比学习更善于从 fMRI 中提取结构清晰、表达稳定的语义特征；而融合模型则在多模态任务中具有更强的整体判别能力。在今后的工作中，可以考虑引入混合式策略，如在融合结构中引入对比学习损失，进一步结合结构建模与目标优化，以实现更强鲁棒性和泛化能力的 fMRI 解码系统。