天津科技大学学报
Journal of Tianjin University of Science & Technology
ISSN 1672-6510,CN 12-1355/N

《天津科技大学学报》网络首发论文

题目：基于上下文掩码与多模态对齐两阶段的脑视觉重建方法
作者：杨巨成，董璇，王嫄，潘旭冉
DOI：10.13364/j.issn.1672-6510.20250001
收稿日期：2025-01-03
网络首发日期：2025-06-23
引用格式：杨巨成，董璇，王嫄，潘旭冉．基于上下文掩码与多模态对齐两阶段的脑视觉重建方法[J/OL]．天津科技大学学报．
https://doi.org/10.13364/j.issn.1672-6510.20250001

2025 年　天津科技大学学报 Journal of Tianjin University of Science and Technology　2025

DOI：10.13364/j.issn.1672-6510.20250001

基于上下文掩码与多模态对齐两阶段的脑视觉重建方法

杨巨成，董　璇，王　媛，潘旭冉
(天津科技大学人工智能学院，天津 300457)

摘　要：脑视觉解码旨在揭示大脑视觉编码机制，从功能性磁共振成像（functional magnetic resonance imaging，fMRI）数据中解码视觉刺激。传统方法重建的视觉刺激常因特征丢失而缺乏语义信息。本文提出一种脑视觉解码框架，该框架融合了上下文掩码与多模态对齐两阶段编码器实现脑视觉重建，以增强视觉刺激的重建质量。首先，双阶段 fMRI 自编码器特征学习模块，其中第一阶段采用 fMRI 上下文掩码自编码器（fMRI contextual mask autoencoder，fCAE）提取图像去噪后的特征表示，并引入潜在上下文回归器以减少特征丢失；第二阶段通过多模态特征对齐进一步优化 fCAE 编码器，以增强重建视觉刺激的语义信息。其次，潜在扩散模型的视觉重建模块，该模块以 fCAE 编码器的输出作为控制条件，实现从大脑活动到视觉刺激的精确重建。实验结果表明，相较于基准模型，本方法在重建视觉刺激的语义准确性（CLIP Score）上提升了 10%。

关键词：脑视觉解码；功能性磁共振成像；深度学习；fMRI 上下文掩码自编码器；多模态对齐；语义重建
中图分类号：TP391　　**文献标志码：**A

Two-Stage Brain Vision Reconstruction Method Based on Contextual Masking and Multimodal Alignment

YANG Jucheng, DONG Xuan, WANG Yuan, PAN Xuran
(College of Artificial Intelligence, Tianjin University of Science and Technology, Tianjin 300457, China)

Abstract: Decoding brain vision aimed to uncover the encoding mechanisms of the brain's visual system within Functional Magnetic Resonance Imaging (fMRI) data to decode visual stimuli. Although traditional methods had been developed to address these issues to some extent, the reconstructed visual stimuli often lacked semantic information due to feature loss. To address this challenge, this paper proposed an innovative framework for brain vision decoding that integrated context masking with a multimodal alignment two-stage encoder to enhance the quality of visual stimulus reconstruction. Firstly, a two-stage fMRI autoencoder feature learning module, where the initial stage employed a fMRI Context-Mask Autoencoder (fCAE) to extract denoised feature representations and introduced a latent context regressor to reduce feature loss; the second stage further optimized the fCAE encoder through multimodal feature alignment to enhance the semantic information of reconstructed visual stimuli. Secondly, a latent diffusion model for visual reconstruction, which utilized the output of the fCAE encoder as a control condition to achieve precise reconstruction from brain activity to visual stimuli. Experimental results indicated that, compared with the baseline methods, this approach improved the semantic accuracy (CLIP Score) of

收稿日期：2025-01-03；**修回日期：**2025-04-13
基金项目：天津市研究生科研创新项目(2022SKYZ370)
作者简介：杨巨成(1980—)，男，湖北天门人，教授，jcyang@tust.edu.cn

reconstructed visual stimuli by 10%.

Key words: brain vision decoding; functional magnetic resonance imaging; deep learning; fMRI Contextual Mask Autoencoder; multimodal alignment; semantic reconstruction

人类视觉系统以其复杂而高效的信息处理能力，在感知和理解外部世界方面扮演着重要的角色，该系统的能力至今仍是人工智能领域追求模拟和复现的目标[1]。近年来，随着脑机接口的不断发展，从大脑活动中解码视觉感知已经成为研究热点[2]。脑视觉解码旨在从收集的大脑信号中提取相应的视觉刺激信息如语义信息。深入解析大脑活动信号并重建相应的视觉刺激是脑视觉解码的任务之一，与简单的语义分类相比，视觉刺激重建对于揭示大脑的复杂工作机制更有意义，也更具挑战性[3]。视觉刺激包含各种图像，其中自然图像的重建是最困难的。通过构建能够准确解码大脑对视觉刺激反应的模型，不仅能够揭示人类视觉系统与计算视觉模型之间的内在联系[4-6]，而且对于辅助那些有交流障碍的病患通过大脑信号传达思想和意愿具有重要意义[7]。

随着图像生成技术的精进和神经影像数据量的增长，研究人员对脑视觉解码的图像重建领域的关注日益增加[8]。fMRI 作为一种非入侵性的大脑活动测量技术，被广范应用于解码脑神经活动。近期研究揭示了人类视觉系统的神经编码和解码过程与深度学习框架之间的相似性[9-11]，这促使研究者广泛利用深度学习网络模型解码神经记录，并重建视觉刺激。

在脑视觉解码的图像重建领域，主要存在两大类模型：优化模型和生成模型。优化模型，例如 Shen 等[12]提出的 DGN 模型，通过将深度神经网络（deep neural network, DNN）提取图像特征作为优化目标，调整图像生成器的潜在空间以匹配解码后的 DNN 特征。然而，由于潜在空间缺乏先验知识，这种从高斯噪声出发的优化过程常常导致结果不明确，且语义信息模糊。相对而言，生成模型，如变分自编码器（variational autoencoder, VAE）[13]、生成对抗网络（generative adversarial network, GAN）[14]和扩散模型（diffusion models, DM）[15]，通过将 fMRI 信号映射到其潜在空间，利用其生成能力，能重建与视觉刺激语义相似的图像，因此备受研究者的青睐。

尽管如此，fMRI 数据中自带的噪声会掩盖与刺激物相关的神经信号，从而增加了从 fMRI 数据直接解码视觉信息的复杂性。此外，通过对大脑活动的研究发现，视觉刺激引起的神经反应过程是复杂的和多阶段的，由此产生的 fMRI 信号是该过程的高度卷积和非线性表示，因此想要逆转这一过程，实现脑视觉解码，重建人脑感知的自然图像是一项极具挑战性的任务。针对噪声问题，研究发现 fMRI 的空间冗余，因此研究者们利用广义的图像去噪自动编码器（掩码自动编码器（masked autoencoders, MAE））进行图像去噪尝试[7,9]，并结合条件潜在扩散模型，实现视觉解码。考虑到大脑机制的复杂性，研究者不再通过简单的岭回归或 L2 线性模型实现 fMRI 体素数据到图像特征的映射，而是通过多层感知机（multilayer perceptron, MLP）或卷积网络进行复杂映射[5,9,16-17]。

虽然上述研究解决了图像去噪和非线性映射的问题，但重建图像仍存在语义模糊的问题，尤其当视觉刺激为自然图像时，该问题格外显著。目前使用的图像去噪模型存在特征丢失问题，从 fMRI 体素数据中提取视觉刺激的语义信息仍不足。为此，本文提出了基于上下文掩码与多模态对齐的脑视觉解码方法。该方法由两大模块组成。首先，双阶段 fMRI 自动编码器的表征学习模块，在第一阶段，fMRI 上下文掩码自编码器（fCAE）通过自监督学习 fMRI 的图像去噪表示，有效避免了 fMRI 数据的空间冗余问题和噪声干扰，同时针对特征丢失问题，在原图像去噪模型 MAE 的基础上添加了潜在上下文回归器，分离了解码器的特征预测和下游任务，在第二阶段丢弃解码器时不会造成特征丢失的问题。由于该模型并非简单的线性模型，在一定程度上模拟了神经编码过程之间的非线性关系。第二阶段的作用则是通过多模态特征对齐，使 fMRI 编码器专注于对图像重建有意义的大脑信号，提高编码质量，增强重建的语义信息。其次，潜在模型的视觉重建模块，以 fCAE 编码器的输出作为控制条件，实现从大脑活动到视觉刺激的精确重建。

1 相关工作

1.1 fMRI 视觉解码

功能性磁共振成像在视觉解码领域的研究自 Haxby 等[18]于 2001 年的开创性工作以来，已经涵盖了分类[19]、识别[20]和重建[4]等多种神经解码任务。视觉刺激的重建对于脑机接口的发展具有重要意义。早期的图像重建技术主要依赖线性回归模型和手工特征提取[21-23]，但这些方法的精确度有限。随着深度学习技术的进步，深度神经网络（DNN）凭借其强大的特征学习能力，显著提升了重建的准确性。Bely 等[24]和 Gaziv 等[25]通过半监督学习[26]训练编码器-解码器结构重建图像，虽解决了刺激图像与 fMRI 配对不足的问题，但重建结果在语义信息的精确度上仍有待提高。Du 等[27-29]进一步引入多视图重建模型，利用 DNN 和高斯先验提高了自然图像和人脸重建的质量。

近期研究趋势显示，深度生成模型，如生成对抗网络和扩散模型，通过将 fMRI 信号映射到图像特征并微调预训练模型，以生成图像。将 fMRI 信号编码到预对齐的视觉-语言潜在空间，并使用预训练的 StyleGAN2 生成图像[30]。Takagi 等[5]开发了一种基于稳定扩散模型的 fMRI 图像重建方法，该方法通过将大脑活动解码成文本描述，并以此作为条件生成自然图像。尽管这些方法较传统回归模型有所改进，但由于直接使用 fMRI 数据进行训练和表征学习，缺乏明确的图像去噪步骤，导致生成图像的质量仍有提升空间。Chen 等[9]采用类似 MAE 的方法对 fMRI 数据进行预训练，通过掩码自编码器自监督学习的方式实现一定程度的图像去噪和特征学习，但未能与图像特征对齐，导致重建图像与刺激图像在语义上存在差异。Sun 等[7]在此基础上引入来自图像自动编码器的像素级指导，帮助区分视觉相关的神经活动与背景噪声，尽管有所优化，但语义信息依然不足。

1.2 掩码模型

掩码模型最初在自然语言处理领域得到广泛应用，随后扩展至计算机视觉领域，在自监督学习中受到广泛关注，并在多种下游任务中展现出卓越的性能。掩码图像建模（masked image modeling，MIM）通过部分遮盖输入图像并从可见部分预测掩码部分，作为预训练任务。预训练得到的编码器能够提取富含语义的信息表征，这些表征随后被应用于下游任务。典型的 MIM 方法，如 BEiT [31]和 iBoT [32]，均采用单一的 ViT（Vision Transformer）架构解决预训练问题，包含学习编码器（表示）和重建掩码补丁两个任务。随后的掩码自动编码器（MAE）将这两个任务部分解耦，目标是从可见部分数据中恢复原始数据[33]。

近期，MAE 因其表征学习和图像去噪能力而被改编应用于 fMRI 解码研究。例如，Chen 等[9]利用 SC-MBM 的稀疏掩码模型模拟大脑中的稀疏编码，并构建了一种生物学上有效的脑特征学习器，用于 fMRI 解码。Sun 等[7]在其基础上提出了双对比掩码自动编码器（DC-MAE），增加来自图像编码器的像素级指导。MAE 结构通常采用编码器-解码器格式，但真正有价值的是学到的编码器，一个高质量的预训练编码器可以显著提升下游任务的性能。Chen 等[34]提出了上下文自动编码器（CAE）方法，通过编码器-回归器-解码器架构将学习编码器与下游任务分离，以提升编码质量。Zhang 等[35]在此基础上利用 CLIP 进行重建监督，进一步提高编码质量。

1.3 扩散模型

扩散模型作为一种先进的生成模型，在图像多样性和保真度方面展现出超越传统生成对抗网络的潜力。这些模型基于参数化的双向马尔科夫链，包括前向扩散过程和反向图像去噪过程。在前向过程中，数据逐步被转化为高斯噪声；而在反向过程中，通过建模后验分布 $p(x)$恢复原始数据，从而生成数据分布的样本[36]。

扩散模型早期主要应用于像素空间，成功生成了高质量图像，但这一过程伴随着推理时间的延长和高昂的训练成本[37]。为了解决这些挑战，潜在扩散模型（latent diffusion model, LDM），也称为稳定扩散（stable diffusion, SD），被提出并应用于图像重建领域[38]。LDM 通过利用预训练的矢量量化生成对抗网络（vector quantized generative adversarial network, VQGAN）或变分自动编码器（VAE）构建高效的潜在图像空间，使优化和评估过程得以在此空间内高效进行[38]。

LDM 不仅能够生成高质量的图像，还能显著减轻计算负担[37]。此外，通过在扩散 U-Net 模型的注意块中引入交叉注意机制，LDM 在图像合成中提供广泛的控制能力，包括文本控制以及对特定领域图

像（如深度图、草图或框架图）的控制[39]。这种多功能性和适应性使得 LDM 在图像合成领域取得了实质性的进步，并拓宽了其应用范围。

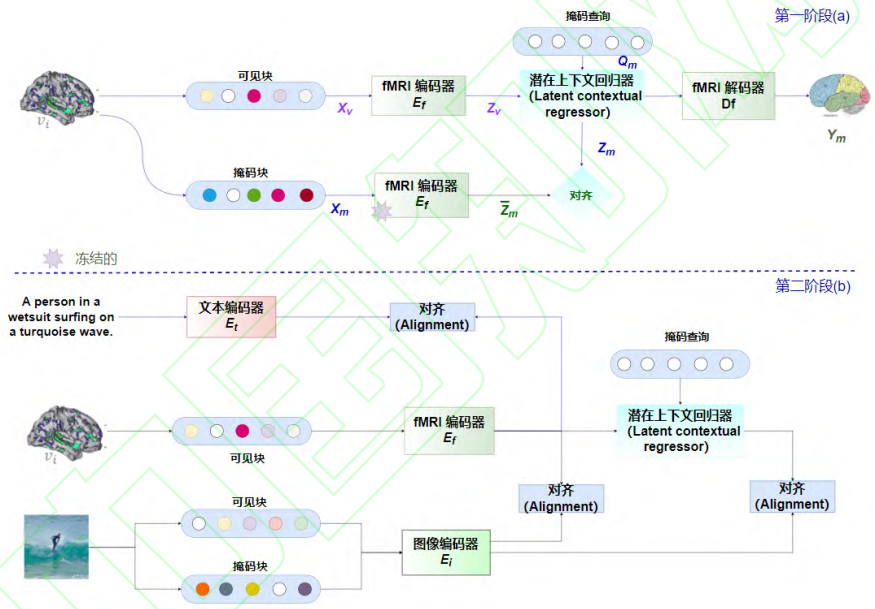
2　本文方法

本文提出的基于上下文掩码和多模态对齐的脑视觉解码方法，由两大模块组成，分别是双阶段 fMRI 自编码器的表征学习和潜在扩散模型的视觉重建。

2.1 双阶段 fMRI 自动编码器的表征学习

2.1.1　基于上下文掩码自编码器的 fMRI 图像去噪表示

双阶段 fMRI 自编码器的表征学习框架如图 1 所示。在第一阶段，本文提出了一种创新的 fMRI 上下文掩码自编码器（fCAE），旨在通过掩码计算消除 fMRI 数据中的噪声和空间冗余，并实现 fMRI 数据的有效表征学习。fMRI 的掩码重建如图 2 所示。由于 fMRI 空间的冗余，即使很大一部分被屏蔽，fMRI 数据仍可以被恢复，且实现了一定程度的图像去噪。该模型的设计灵感源自掩码图像建模方法[7,9]。与先前工作所使用的 MAE 架构相比，本阶段的 fCAE 额外添加了 1 个潜在上下文回归器，主要功能是在编码的表示空间中对掩码块进行预测，用来预训练编码器，使解码器只专注于下游任务，而不参与掩码表征预测，以此提高编码器的编码质量。



第一阶段（a）为 fMRI 上下文掩码自编码器，第二阶段（b）为多模态对齐的 fMRI 表征学习。

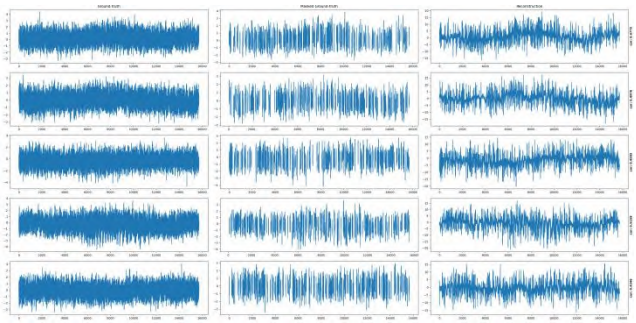
图 1　双阶段 fMRI 自编码器的表征学习框架

Fig. 1　Two-stage fMRI autoencoder representation learning framework

本阶段的预训练模型如图 1(a)所示，模型架构主要由 fMRI 编码器、潜在上下文回归器和 fMRI 解码器三部分组成。fMRI 编码器负责将可见块和掩码块编码成特征表示；潜在上下文回归器则基于可见块的编码特征 Z_v 和掩码查询 Q_m 预测掩码部分的表征 Z_m；fMRI 解码器则利用预测的 Z_m 重建掩码部分。该模型的核心优势在于通过引入潜在上下文回归器和表示对齐机制，潜在上下文回归器能够生成掩码块的语义表示代替 MAE 中解码器的预测掩码块的表示，使该模型的解码器专注于重建任务，编码器能够专注于学习高质量语义表示，而不会被解码器干扰，即使最后丢弃解码器也不会造成特征的丢失。而潜在上下文回归器的掩码块预测与编码器可见块的表示进行对齐，确保两者表示位于同一语义空间中，从而为重建任务提供了丰富的语义信息。。

在实验中，本文将处理后的 fMRI 图像随机分割为可见块 X_v 和掩码块 X_m。在编码环节，采用 ViT 架构构建编码器 E_f。通过多层感知机和一维卷积层将 fMRI 数据转换成高位嵌入形式，以适配模型的输入需求。在此基础上，引入位置嵌入（positional

embeddings）和分类标记(Cls_Token)，并与可见块 X_v 结合，以捕捉序列中元素的顺序信息和全局上下文。这些嵌入随后被送入到自注意力（self-attention）模块，以生成可见块的潜在表征 Z_v。



（a）真实 fMRI 数据（b）掩码后 fMRI 数据（c）重建 fMRI 数据

图 2 fMRI 的掩码重建

Fig. 2 Mask reconstruction of fMRI

在预测环节，潜在上下文回归器基于编码器输出的可见块潜在表征 Z_v，预测掩码块的潜在表征 Z_m。在此过程中，特别考虑掩码块的位置信息，以增强模型对空间结构的敏感性。为此，引入了交叉注意力（cross-attention）模块，这是潜在上下文回归器的主要结构，该模块允许模型将可见块的信息 Z_v 与掩码块的位置信息相结合，以预测掩码块的潜在表征 Z_m。掩码查询 Q_m，作为交叉注意力模块中的掩码标记，通过学习过程中的反向传播进行优化。

在解码环节，采用与编码器相似的架构，由多个自注意力模块堆叠而成，最后是预测目标的线性层。解码器只负责将预测的潜在表征 Z_m 重建为掩码块 Y_m，这种策略允许解码器专注于重建任务。

关于损失函数设计，采用重建损失 ℓ_{rec} 和对齐损失 ℓ_{ali}。在重建损失中，Y_m 是解码器输出的预测结果，$\bar{Y}_m$ 是掩码块的离散化表示。在对齐损失中，将掩码块 X_m 送入冻结的编码器 E_f 生成连续掩码表征 $\bar{Z}_m$ 作为目标，使其与连续 Z_m 表征进行对齐重建损失确保掩码部分的重建质量，而对齐损失则确保掩码表示的预测精度。这两个损失函数的加权和构成了模型的总损失，为

$$\ell_{\text{rec}}\left(Y_m,\bar{Y}_m\right)+\lambda\ell_{ali}\left(Z_m,sg[\bar{Z}_m]\right) \qquad (1)$$

其中：ℓ_{rec} 用交叉熵损失函数实现，交叉熵损失函数能够对离散化的目标有效地衡量预测的准确性和置信度，ℓ_{ali} 使用的是均方误差损失函数，均方误差损失能够对连续的表示有效地衡量预测的准确性和误差大小。sg[·] 代表停止梯度传播，即在计算对齐损失时，不更新的 $\bar{Z}_m$ 梯度，λ 代表损失权重参数。

2.1.2　基于多模态对齐的 fMRI 表征学习

第二阶段，如图 1（b）所示，对比第一阶段模型，舍弃了 fMRI 解码器组件，专注于对 fMRI 编码器进行优化。鉴于掩码模型的目标特性，本阶段选择将语义丰富的图像特征和文本嵌入作为目标，实现多模态的对齐，进一步优化 fMRI 编码器，使其能够学习由目标带来的丰富语义信息，增强表征学习能力。本阶段采用编码器-潜在上下文回归器架构。

首先，在 fMRI 编码器部分，E_f 只接收可见块 X_v，并通过 VIT 模型（ViT-Base）得到潜在表征 Z_v。随后，与第一阶段相同，使用潜在上下文回归器预测掩码块的潜在表征 Z_m。初始查询是可学习的掩码标记 Q_m。

以两个 CLIP 潜在特征为目标与 fMRI 潜在表征进行特征对齐，其中目标特征中的图像特征分 I_v 和 I_m，文本特征为 T。损失函数由两部分组成：fMRI 的潜在表征与图像的潜在表征对齐 $\ell_{\cos}$，包括可见和掩码部分；fMRI 的潜在表征与文本嵌入对齐 ℓ_{con}。损失函数为

$$\ell_{\cos}\left\{cat\left(Z_v,Z_m\right),\ cat\left(I_v,I_m\right)\right\}+\ell_{con}\left(Z,T\right) \qquad (2)$$

在实验中，对齐的损失函数分别用的是余弦距离损失和对比损失函数。余弦距离损失为

$$\ell_{\cos}=\frac{1}{n}\sum_{i-1}^{|n|}(1-\cos(y^i-\bar{y}^i))$$
$$\cos(y^i-\bar{y}^i)=\frac{y^i\cdot\bar{y}^i}{\left\|y^i\right\|\left\|\bar{y}^i\right\|} \qquad (3)$$

其中：$y^i=cat(Z_{\text{v}},Z_{\text{m}})$，$\bar{y}^i=cat(I_{\text{v}},I_{\text{m}})$，$y^i$ 表示第 i 个 fMRI 潜在特征，$\bar{y}^i$ 表示第 i 个图像特征，cat 特征连接是沿着第一维度进行连接。余弦距离损失函数衡量了 fMRI 潜在特征与图像特征之间的相似度，通过最小化这一损失，能够促使 fMRI 编码器学习到与图像特征更为接近的表征。

2.2 潜在扩散模型的视觉重建

在经过第一阶段和第二阶段的训练后，以优化后的 fMRI 编码器输出为 LDM 提供条件控制信号，

从而指导 LDM 从大脑活动中重建图像。潜在扩散模型的视觉重建如图 3 所示。扩散模型由前向扩散过程和后向图像去噪过程组成。正向过程通过逐步添加噪声，每一步添加的噪声从标准正态分布 $N(0,1)$ 中采样，每一步 t，$\varepsilon_t \sim N(0,1)$ 被用来更新 z_t，该过程可以表示为

$$\mathrm{z}_t = \sqrt{\overline{\alpha}_t}\mathrm{z}_0 + \sqrt{1-\overline{\alpha}_t}\varepsilon_t \qquad (4)$$

其中：z_0 是 LDM 的初始潜在变量，z_t 是在时间步长 t 的 LDM 潜在变量，$\varepsilon_t \sim N(0,I)$ 是标准正态分布噪声，$\overline{\alpha}_t = \prod_{i=0}^{t}\alpha_i$，$\alpha_t$ 是一个经验常量，随着 t 的增大而减小。反向图像去噪过程的目标是从噪声 z_t 中逐步恢复出原始数据 z_0。时间步 t 是从 $[0,\ T-1]$ 中随机采样的，T 为扩散步数。这个过程由神经网络（通常是 U-Net）实现，网络的任务是预测每一步添加的噪声 ε_t。

在训练阶段，扩散模型，尤其是 U-Net 模型，学习在前向过程中逐步添加噪声，并在后向过程中通过消除这些噪声生成图像。潜在扩散模型在低维潜在空间中执行这一过程，该空间通过预训练的 VQ-VAE 或 VQ-GAN 模型构建，有效降低了计算复杂度，同时保持了图像生成的质量。本文将 LDM 作为图像生成模型的骨干，并结合上述训练的 fMRI 自编码器作为条件控制模型，通过基于注意力的 U-Net 中的交叉注意力机制实现条件控制。

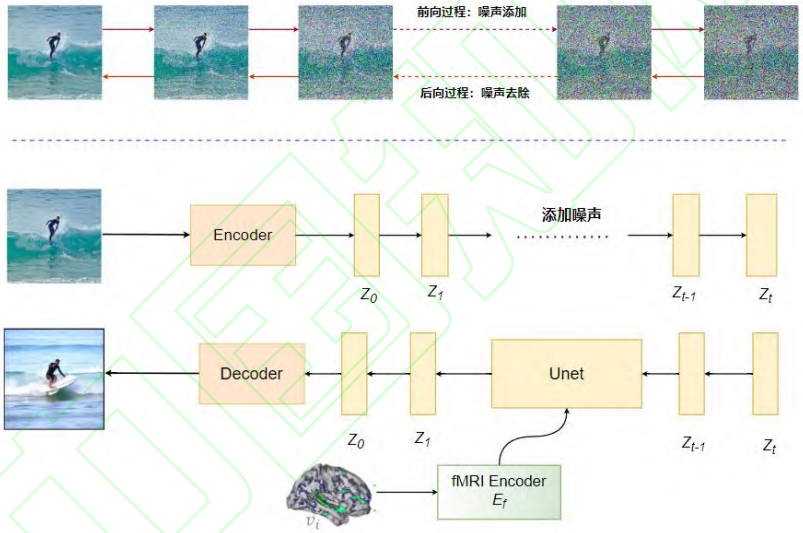


图 3 潜在扩散模型的视觉重建

Fig. 3　Latent diffusion models for visual reconstruction

将 fMRI 解码为自然图像的任务可以视为有条件的图像生成问题。考虑到 fMRI 的信噪比较低，且 fMRI 到图像的数据对数量有限，直接训练 fMRI 到图像的生成模型面临巨大挑战。因此，本阶段的目标是利用 fMRI 从预训练的条件图像生成模型中提取与图像相关的信息，而本文所采用的模型是预训练的文本到图像的 LDM，并在 fMRI 的引导下生成图像。在多数图像生成任务中，对不同模式的多样性进行采样至关重要，例如标签到图像和文本到图像的转换。然而，fMRI 到图像的转换更侧重于生成的一致性，预期从大脑活动解码的图像在语义上相似。因此，需要一个更强大的调节机制确保这种一致性，尤其是在概率扩散模型中。为此，本文将交叉注意力调节与时间步骤调节相结合，为任务提供更精确的指导。在训练中，给定一个图像 I 和 fMRI 数据 f 以及 VQ-GAN 编码器 E_g 和 fMRI 编码器 E_f，并冻结 LDM 使用以下损失（式（5））对 fMRI 编码器进行微调。

$$L_{\mathrm{t}} = E_{t,u_0,\varepsilon_t\sim N(0,1)}\left[\left\|\varepsilon_t - \varepsilon_\theta\left(\phi\left(E_{\mathrm{g}}(I)\right)_t, t, E_f(f)\right)\right\|_2^2\right] \qquad (5)$$

其中：t 从 T 到 1，$E_{t,I,\varepsilon\sim N(0,1)}$ 表示对时间 t，原始图像 I 和噪声 ε_t 的期望，ε_θ 为模型预测的噪声，其中 θ

表示模型的参数 $\phi\left(E_g(I)\right)_t=\sqrt{\overline{\alpha}_t}E_g(I)+\sqrt{1-\overline{\alpha}_t}\varepsilon_t$ 是经过训练的 VQGAN 编码器 E_g 编码后的潜在表示，并且在时间步 t 被噪声化，$E_f(f)$ 为经过训练的 fMRI 编码器的 fMRI 数据表示。该损失函数用于优化模型，让模型学习在给定 fMRI 编码结果的条件下预测在扩散过程中添加的噪声。在推理阶段，该过程从时间步长 t 的标准高斯噪声开始。LDM 依次进行向后过程，逐步对潜在表征进行图像去噪，以给定的 fMRI 信息为条件。在时间步长为 0 时，VQGAN 解码器被用来将潜在的表征转换为图像。

3 实　验

3.1 数据集

NSD（natural scenes dataset）[40]是一个大规模的 fMRI 数据集，旨在促进认知神经科学与人工智能的交叉融合。该数据集通过提供丰富的自然场景刺激，为深入研究大脑活动提供了宝贵的资源。本研究记录了 8 名受试者的 BOLD 响应，受试者观看的图像刺激源自 Common Objects in Context（COCO）[41]数据集，图像刺激的文本描述可以通过 COCO ID 检索。由于本数据集数据庞大且图像刺激内容丰富，故仅使用该数据集进行模型训练。

为评估实验在不同受试者间的稳定性，本研究选取了 NSD 中编号为 1、2、5 和 7 的受试者的 fMRI 和图像刺激信息，COCO ID 数据，处理后构成 fMRI-图像-COCO ID 数据集。后依据 COCO ID 获取了每个图像的 1~3 条文本描述，并根据 COCO ID 的坐标按顺序单独存放到文件中，使用时取文本特征的平均值。每位受试者的训练集包含 8192 个三模态样本，测试集则包含 900 个样本。测试集中的图像刺激是 4 位受试者共享的。具体的样本分布详见表 1。

表 1　NSD 数据集

Tab. 1　NSD of dataset

数据集	训练	测试	受试者 ID	体素
NSD	8192	900	Sub01	15725
			Sub02	10285
			Sub05	9551
			Sub07	9187

3.2 实验设计

3.2.1 双阶段 fMRI 自动编码器的表示学习

本研究提出的 fCAE 基于 ViT 架构，与传统的掩码自动编码器（MAE）有所不同，采用非对称结构，其中解码器的规模远小于编码器。此外，本文不依赖嵌入-补丁尺寸比，而是采用线性卷积操作，其中卷积维度为 512，输出通道数为 1024。为了增强模型的泛化能力，本文引入了随机稀疏化（random sparsity, RS）技术作为数据增强手段。通过随机选择并设置 fMRI 数据中 20%的体素值为 0，RS 技术能够有效地模拟大脑活动的自然变异性，从而提高模型对异常值的鲁棒性。

fMRI 编码器的训练过程分为两个阶段。在第一阶段，通过均方误差(MSE)损失函数对潜在上下文回归器预测的掩码特征与初始掩码部分的编码特征进行对齐，以确保模型能够有效地学习掩码信息。在此阶段，关于损失函数中的权重，并没有做广泛的研究，只尝试了 3 种选择（λ=1, λ=1.5, λ=2），线性探测的结果分别为 63.4、63.7、64.1，故损失权重被设定为 2，掩码率为 0.75，批次大小为 8，模型经过 100 个 epochs 的训练。所有实验均在 NVIDIA GeForce RTX 3090 GPU 上进行。为了优化训练过程，采用预热策略并设置初始学习率为 1.7×10^{-3}，最小学习率为 1×10^{-4}，同时配合余弦衰减学习率调度器。优化器选用了 AdamW，权重衰减参数设为 0.05。该阶段的参数量为 277 M。第二阶段，进行多模态的对齐优化 fMRI 编码器，其中图像和文本编码器采用 CLIP 模型，其嵌入维度为 512，与 fMRI 编码后的特征维度相匹配。在这一阶段，批次大小增加至 16，训练周期延长至 200 个 epochs。初始学习率调整为 5×10^{4}，优化器继续使用 AdamW，权重衰减参数维持在 0.05。在此阶段，仅对 fMRI 编码器进行调整，而将其他部分冻结，以确保模型的稳定性。整个训练过程大约耗时 13 h，参数量为 185 M。

3.2.2 微调潜在扩散模型

在微调潜在扩散模型的过程中，采取联合优化策略，针对 LDM 的交叉注意力头和 fMRI 编码器的参数进行调整，而保持 LDM 的其他参数固定不变。给定 1 对 fMRI-图像，首先利用预训练的 VQ 编码器（f4-vq）对图像进行编码，获得其潜在表征，该表征将作为后续联合训练的目标指导。在训练过程中，fMRI 数据通过先前阶段训练得到的 fMRI 编码器生

成潜在表示，随后这一表示被进一步投影至 LDM 的交叉注意力模块，并与时间嵌入相结合，实现双重调节。本文遵循扩散模型的训练设置，优化模型以预测在每个时间步长中添加至图像潜在特征的高斯噪声，采用 fMRI 编码器的输出作为调节信息。训练参数设置如下：批次大小为 16，扩散步骤设定为 250，选用 AdamW 优化器，学习率设定为 5.3×10^{-5}，图像分辨率为 256×256×3。

3.3　基线模型和评价指标

为了验证实验的有效性，使用均方误差（MSE）、像素级相关性（PixCorr）、结构相似性指数（SSIM）和 CLIP Score 这 4 个评价指标与基线模型 vis-dec[7] 和 mind-vis[9]进行比较，衡量结构和语义。

MSE 衡量原始图像与失真图像之间的像素值差的平方的平均值，反映图像失真程度。PixCorr 计算重建图像和真实图像在像素级别上的皮尔森相关系数，评估图像重建的质量。SSIM 考虑图像的亮度、对比度和结构 3 个维度，衡量两幅图像的相似度。CLIP Score 基于自然语言描述的图像评估指标，通过比较重建图像和原始图像在 CLIP 模型中的特征相似性来评估图像的语义准确性。

4　结果与分析

4.1　重构结果

基准模型对比见表 2。与另外两个模型相比，本模型在整体上有所提升。重建结果如图 4 所示。在复杂场景下，本文模型的重建质量显著优于其他模型，并且与真实图像的语义一致性更高。

表 2 基准模型对比

Tab. 2　Comparison of benchmark models

模型	参数量	MSE↓	PixCorr↑	SSIM↑	CLIP↑
Mind-vis[9]	303M	0.513	0.214	0.260	0.55
Vis-dec[7]	317M	0.486	0.202	0.251	0.63
本文模型	**277M**	**0.485**	**0.235**	**0.302**	**0.72**

为了检验本文模型是否能可靠地重建不同受试者的大脑活动，将在只训练 sub01 数据集的情况下，同时在 4 个受试者的测试集进行评估和在训练 4 个受试者数据集的情况下，对 4 个受试者的测试集进行评估。图 5 的重建结果显示，本文的模型能够良好的重建不同受试者的大脑活动。

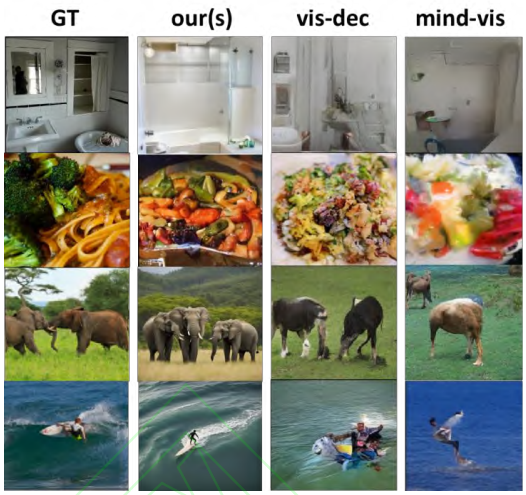
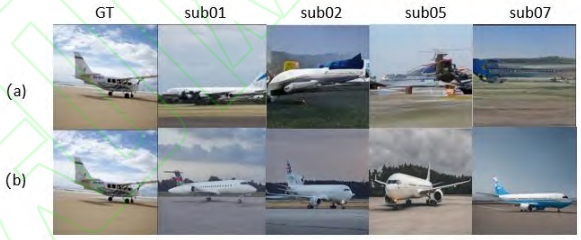


图 4 重建结果

Fig. 4　Reconstruction results



（a）只训练单个受试者，（b）同时训练四个受试者。

图 5 不同受试者的重建结果

Fig. 5　Reconstruction results for different subjects

4.2　消融实验

通过一系列消融实验研究了模型架构不同组件的有效性，特别是编码器设计和多模态信息的使用，图像和文本的编码器是预训练的 CLIP 编码器，目标进行相应特征的提取。不同构造上的消融实验结果的定量比较结果见表 3。表 3 中参数量指的是模型参数量，fMRI 编码器为 fCAE 时，模型参数包括编码器和潜在上下文回归器，fMRI 编码器为 MAE 时，模型参数仅包含编码器。结果表明，本文模型的整体设计显著提高性能，去除某些组件会导致性能明显下降。消除第二阶段图像部分（I）、文本部分（T）或仅保留 fMRI 编码部分（F）都会导致 top-1 精度的降低。这些发现强调了每种模态和组件在模型准确重建大脑活动和视觉刺激能力中的重要性。通过比较完整模型和消融版本的性能，证实图像和文本模态以及 fMRI 编码模型在增强模型从大脑活动中解码视觉信息的能力方面起着至关重要的作用。

表 3 不同构造上的消融实验结果的定量比较

Tab. 3　Quantitative comparison of ablation experiment results on different architectures.

模型	fMRI 编码器	参数量	模态	top-1 准确率
full	**fCAE**	**102M**	**F+I+T**	**26.3%**
1	fCAE	102M	F	17.5%
2	fCAE	102M	F+I	23.4%
3	fCAE	102M	F+T	20.1%
4	MAE	86M	F+I	19.8%
5	MAE	86M	F	16.4%

5　结　语

本研究提出基于上下文掩码与多模态对齐脑视觉解码方法，与现有基线模型相比，本文方法能够更准确地从 fMRI 数据中解码视觉信息，生成与视觉刺激语义相似的图像，从而验证了方法的有效性。本文方法虽有一定程度的提升，但在实际的脑机接口场景中仍面临一些挑战和局限性。首先，计算复杂度和资源需求较高是实际应用中的一个重要的问题。本文模型需要一定的计算资源和较长的训练时间，在实时性要求较高的脑机接口中，例如辅助交流障碍患者快速传达思想，这种高计算复杂度可能会限制模型的实时性响应能力。其次，fMRI 数据的获取和处理也存在一定的局限性。fMRI 数据虽然能够提供丰富的神经活动信息，但其采集过程相对复杂且成本较高，其高噪声差异和个体差异性也可能对模型的泛化能力和鲁棒性产生影响。未来的研究方向可以探索更高效的模型架构，降低计算复杂度，提高模型的泛化能力和鲁棒性。

参考文献：

[1] BELLIVEAU J W, KENNEDY D N, MCKINSTRY R C, et al. Functional mapping of the human visual cortex by magnetic resonance imaging[J]. Science, 1991, 254(5032): 716-719.

[2] LAN Y T, REN K, WANG Y, et al. Seeing through the brain: image reconstruction of visual perception from human brain signals[EB/OL]. [2025-01-01]. https://doi.org/10.48550/arXiv.2308.02510.

[3] 杜长德,何晖光. 基于视觉信息编解码的深度学习类脑机制研究[J]. 张江科技评论, 2019(4):25-27.

[4] REN Z Q, LI J, XUE X T, et al. Reconstructing seen image from brain activity by visually-guided cognitive representation and adversarial learning[J]. Neuroimage, 2021, 228: 117602.

[5] TAKAGI Y, NISHIMOTO S. High-resolution image reconstruction with latent diffusion models from human brain activity[C]//IEEE. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2023: 14453-14463.

[6] SUN J, LI M, MOENS M F. Decoding realistic images from brain activity with contrastive self-supervision and latent diffusion[EB/OL]. [2025-01-01]. https://doi.org/10.48550/arXiv.2310.00318.

[7] SUN J, LI M, CHEN Z, et al. Contrast, attend and diffuse to decode high-resolution images from brain activities[J]. Advances in neural information processing systems, 2023, 36: 12332-12348.

[8] LU Y, DU C, ZHOU Q, et al. Minddiffuser: controlled image reconstruction from human brain activity with semantic and structural diffusion[C]//ACM. Proceedings of the 31st ACM International Conference on Multimedia. New York：ACM, 2023: 5899-5908.

[9] CHEN Z, QING J, XIANG T, et al. Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding[C]//IEEE. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York ： IEEE, 2023: 22710-22720.

[10] PINTO N, DOUKHAN D, DICARLO J J, et al. A high-throughput screening approach to discovering good forms of biologically inspired visual representation[J]. PLOS Computational biology, 2009, 5(11): e1000579.

[11] SCHRIMPF M, KUBILIUS J, HONG H, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? [EB/OL]. [2025-01-01]. https://doi.org/10.1101/407007.

[12] SHEN G, HORIKAWA T, MAJIMA K, et al. Deep image reconstruction from human brain activity[J]. PLOS Computational biology, 2019, 15(1): e1006633.

[13] CHEN Y, LIU J, PENG L, et al. Auto-encoding variational Bayes[J]. Cambridge explorations in arts and sciences, 2024, 2(1): 33.

[14] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[15] SONG Y, ERMON S. Generative modeling by estimating gradients of the data distribution[EB/OL]. [2025-01-01]. https://doi.org/10.48550/arXiv.1907.05600.

[16] MOZAFARI M, REDDY L, VANRULLEN R. Reconstructing natural scenes from fMRI patterns using BigBiGAN[C]//IEEE. 2020 International joint conference on neural networks (IJCNN). New York: IEEE, 2020: 1-8.

[17] CHEN Z, QING J, XIANG T, et al. Seeing beyond the brain: masked modeling conditioned diffusion model for human vision decoding [EB/OL]. [2025-01-01]. https://doi.org/10.48550/arXiv.2211.06956.

[18] HAXBY J V, GOBBINI M I, FUREY M L, et al. Distributed and overlapping representations of faces and objects in ventral temporal cortex[J]. Science, 2001, 293(5539): 2425-2430.

[19] DU C, FU K, LI J, et al. Decoding visual neural representations by multimodal learning of brain-visual-linguistic features[J]. IEEE Transactions on pattern analysis and machine intelligence, 2023, 45(9): 10760-10777.

[20] HORIKAWA T, KAMITANI Y. Generic decoding of seen and imagined objects using hierarchical visual features[J]. Nature communications, 2017, 8(1): 15037.

[21] NASELARIS T, PRENGER R J, KAY K N, et al. Bayesian reconstruction of natural images from human brain activity[J]. Neuron, 2009, 63(6): 902-915.

[22] KAY K N, NASELARIS T, PRENGER R J, et al. Identifying natural images from human brain activity[J]. Nature, 2008, 452(7185): 352-355.

[23] FUJIWARA Y, MIYAWAKI Y, KAMITANI Y. Modular encoding and decoding models derived from Bayesian canonical correlation analysis[J]. Neural computation, 2013, 25(4): 979-1005.

[24] BELIY R, GAZIV G, HOOGI A, et al. From voxels to pixels and back: self-supervision in natural-image reconstruction from fMRI[J]. Proceedings of the 33rd international conference on neural information processing systems,2019,585:6517-6527.

[25] GAZIV G, BELIY R, GRANOT N, et al. Self-supervised natural image reconstruction and large-scale semantic classification from brain activity[J]. Neuroimage, 2022, 254: 119121.

[26] LEARNING S S. Semi-supervised learning[EB/OL]. [2025-01-01]. https://home.ttic.edu/~avrim/MLT18/SSL.pdf.

[27] DU C, DU C, HUANG L, et al. Reconstructing perceived images from human brain activities with Bayesian deep multiview learning[J]. IEEE Transactions on neural networks and learning systems, 2018, 30(8): 2310-2323.

[28] DU C, DU C, HUANG L, et al. Conditional generative neural decoding with structured CNN feature prediction[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(3): 2629-2636.

[29] DU C, DU C, HUANG L, et al. Structured neural decoding with multitask transfer learning of deep neural network representations[J]. IEEE Transactions on neural networks and learning systems, 2020, 33(2): 600-614.

[30] LIN S, SPRAGUE T, SINGH A K. Mind reader: reconstructing complex images from brain activities[J]. Advances in neural information processing systems, 2022, 35: 29624-29636.

[31] BAO H, DONG L, PIAO S, et al. BEiT: BERT pre-training of image transformers[EB/OL]. [2025-01-01]. https://doi.org/10.48550/arXiv.2106.08254.

[32] ZHOU J, WEI C, WANG H, et al. iBOT: image BERT pre-training with online tokenizer[EB/OL]. [2025-01-01]. https://doi.org/10.48550/arXiv.2111.07832.

[33] CAO S, XU P, CLIFTON D A. How to understand masked autoencoders[EB/OL]. [2025-01-01]. https://doi.org/10.48550/arXiv.2202.03670.

[34] CHEN X, DING M, WANG X, et al. Context autoencoder for self-supervised representation learning[J]. International journal of computer vision, 2024, 132(1): 208-223.

[35] ZHANG X, CHEN J, YUAN J, et al. CAE v2: context autoencoder with CLIP target[EB/OL]. [2025-01-01]. https://doi.org/10.48550/arXiv.2211.09799.

[36] YANG L, ZHANG Z, SONG Y, et al. Diffusion models: a comprehensive survey of methods and applications[J]. ACM Computing surveys, 2023, 56(4): 1-39.

[37] DHARIWAL P, NICHOL A. Diffusion models beat GANs on image synthesis[J]. Advances in neural information

processing systems, 2021, 34: 8780-8794.

[38] JIN Z, SHEN X, LI B, et al. Training-free diffusion model adaptation for variable-sized text-to-image synthesis[J]. Advances in neural information processing systems, 2023, 36: 70847-70860.

[39] LIU B, WANG C, CAO T, et al. Towards understanding cross and self-attention in stable diffusion for text-guided image editing[C]//IEEE. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York：IEEE, 2024: 7817-7826.

[40] ALLEN E J, ST-YVES G, WU Y, et al. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence[J]. Nature neuroscience, 2022, 25(1): 116-126.

[41] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco: Common objects in context[C]// ECCV. Computer Vision-ECCV 2014: 13th European Conference. Berlin：Springer International Publishing, 2014: 740-755.