利用层次化视觉嵌入学习大脑表征

Jiawen Zheng[1]*、**Haonan Jia**[2]*、**Ming Li**[3]、**Yuhui Zheng**[1]、**Yufeng Zeng**[1]、**Yang Gao**[2]、**Chen Liang**[1] †
[1]香港科技大学（广州）[2]北京航空航天大学 [3]清华大学
jzhengbx@connect.hkust-gz.edu.cn, chenliang2@hkust-gz.edu.cn

arXiv:2602.07495v1 [cs.CV] 7 Feb 2026

摘要

从脑信号中解码视觉表征，已成为神经科学和人工智能领域广受关注的研究方向。然而，脑信号在多大程度上真正编码了视觉信息，仍不清楚。现有视觉解码方法探索了多种脑—图像对齐策略，但大多侧重于高层语义特征，忽视像素级细节，从而限制了我们对人类视觉系统的理解。本文提出一种脑—图像对齐策略，利用具有不同归纳偏置的多个预训练视觉编码器来捕获层次化与多尺度视觉表征，并采用对比学习目标，实现脑信号与视觉嵌入之间的有效对齐。此外，我们引入融合先验，在大规模视觉数据上学习稳定的映射，随后将脑特征与这一预训练先验进行匹配，从而增强不同模态间的分布一致性。大量定量和定性实验表明，我们的方法在检索准确率与重建保真度之间取得了良好平衡。

1 引言

随着文生图生成模型的快速发展（Rombach 等人，2022；Zhang 等，2023；Esser 等人，2024），从脑信号中重建人类视觉刺激已成为神经科学和人工智能领域备受关注的研究方向。视觉处理是人脑的核心功能。当视觉刺激被大脑处理时，初级视觉皮层首先解析颜色、边缘和纹理等基本像素属性，随后将其传递至不同的高阶视觉皮层，进行进一步的层级处理（Blasdel & Lund，1983；Tsumoto 等人，1978）。这些高阶区域协同整合并概括视觉数据，形成物体和环境等语义特征，从而构成人类视觉感知外部世界的基本过程。（Merigan 和 Maunsell，1993）。

为探究人类视觉系统与脑表征之间复杂且动态的关系，研究人员通常采用功能性磁共振成像（fMRI）、脑磁图（MEG）和脑电图（EEG）进行视觉解码与重建（Zhang 等，2025；Benchetrit 等人，2023）。功能性磁共振成像通过血氧水平依赖信号间接测量大脑活动，具有较高的空间分辨率，但时间分辨率有限，因此难以捕捉快速的神经动力学变化（Logothetis 等人，2001）。相比之下，脑电图和脑磁图直接反映大脑的电生理活动。脑电图具有较高的时间分辨率，但空间分辨率较低且信噪比较差。脑磁图同样具有毫秒级时间精度，其空间分辨率则相对更高（Liu 等人，2023a；da Silva，2013）。

先前研究通过将脑信号与视觉表征对齐，探索脑信号解码，从而实现分类、检索和重建。Song 等人（2024）采用对比学习，在匹配的脑信号–图像对之间最大化相似度，同时最小化不匹配

*技术贡献相同。
†通讯作者。

对之间的相似度。Li 等人（2024）提出了自适应思维映射器（ATM），将脑信号特征与 CLIP 派生的视觉嵌入对齐，并结合两阶段多管道策略实现从脑信号到图像的生成。然而，这些方法依赖于脑信号与图像特征之间的直接对齐，而两种模态之间的结构差异使得这种策略不足以捕捉潜在的共享表征。
近期，一些研究尝试通过引入先验或丰富视觉表征来改善直接对齐。Wu 等人（2025）提出了不确定性感知模糊先验（UBP），通过模糊图像高频细节来缓解脑图像不匹配。Zhang 等人（2025）将深度信息融入 CLIP 衍生的图像嵌入，以增强脑—图像对齐。然而，这些方法主要关注高层语义对齐，却忽略了低层像素信息。这一疏忽妨碍了人们全面理解脑信号所编码的视觉内容，并降低了可解释性。

为弥合脑信号时间动态与视觉表征层次结构之间的结构差异，我们受人类视觉系统感知机制的启发，提出了一种带有融合先验的层次化视觉融合框架。该框架整合多个预训练编码器，构建从像素级细节到高层语义的多尺度视觉表征，并利用对比学习对齐脑特征与视觉特征。为弥补 CLIP 及相关编码器在捕捉局部和细粒度信息方面的局限性，我们将由变分自编码器（VAE）建模的低层视觉特征融入融合表征。这种融合显著提升了零样本检索性能。此外，我们在大规模视觉数据上预训练融合先验，以提供从融合特征到扩散条件的稳定映射。将脑嵌入与该先验对齐，可实现忠实的图像重建，并作为脑到图像解码框架的有效桥梁。

- 我们提出了一种基于融合的脑视觉接口，将脑嵌入与由层次化编码器（语义和像素）构建的融合视觉词元对齐，并结合预训练的融合先验，将该词元连接到冻结的图像生成骨干网络，实现可复用且无需文本的生成。
- 我们得出了一项关键科学发现：脑电/脑磁信号能够同时与高层语义和低层视觉细节对齐。在我们的融合框架中，在语义编码器之上添加 VAE 潜变量能够持续提升解码性能，而仅使用语义或仅使用像素的设置都无法恢复相同的脑视觉结构。
- 我们的方法实现了最先进的零样本检索性能，并提升了重建质量；同时，在固定的基于融合的训练方案下，该方法可即插即用于不同的脑编码器。

2 相关工作

脑视觉解码 神经解码旨在从脑电图（Song et al., 2024; Bai et al., 2023; Li et al., 2024）、脑磁图（Cichy et al., 2016b）或功能性磁共振成像（Kay et al., 2008; Takagi & Nishimoto, 2023b）等脑信号中推断人类的认知与知觉状态。其中，视觉解码已成为一项尤具挑战性且前景广阔的研究方向，主要涵盖图像分类（Xu et al., 2024）、检索（Liu et al., 2023c）和重建（Ozcelik & Van- Rullen, 2023; Takagi & Nishimoto, 2023a）等任务。研究的核心重点之一，是将脑电信号编码为有效的表征向量，以捕捉其时间与频率特征（Fu et al., 2025）。为弥合模态差异，通常采用基于 CLIP 的模型作为基准方法（Liu et al., 2023c; Wang et al., 2024）；近期方法则探索了多种策略，例如利用扩散先验（Aggarwal et al., 2023）增强生成空间中的语义一致性（Ozcelik & VanRullen, 2023; Takagi & Nishimoto, 2023b; Li et al., 2025），或通过双向映射来确保跨模态循环一致性（Wei et al., 2024）。与此同时，针对图像模态本身的研究也探索了多种方法，以弥补神经信号语义表达能力有限的问题，包括采用模糊预处理来抑制高频噪声（Li et al., 2024），以及利用文本描述丰富语义引导（Takagi & Nishimoto, 2023b）。然而，现有方法大多侧重于高层语义，未能充分捕捉像素级的精细表征，因此生成或重建图像的保真度仍存在明显不足。

层次化与多尺度视觉表征 图像单模态表征学习的最新进展，强调在单一模态内学习多层次语义与密集结构。通过自监督训练的视觉Transformer（Dosovitskiy et al., 2020）无需文本监督即可获得强大的全局语义表征（Bao et al., 2021; Xie et al., 2022; He et al., 2022; Caron et al., 2021; Oquab et al., 2023）；而变分自编码器和矢量量化变分自编码器等生成式潜变量模型，则能提供紧凑的像素级编码，并实现高重建保真度（Kingma & Welling, 2013; Higgins et al., 2017; Van Den Oord et al., 2017; Razavi et al., 2019）。神经科学的另一项相关发现表明，较深层的网络特征与高级视觉区域相关，而较早的层则对应精细空间细节和快速动态（Yamins et al., 2014; Cichy et al., 2016a），这促使研究者将粗粒度的语义抽象与细粒度的局部线索相结合。然而在实践中，许多解码流程（Li et al., 2024; Wu et al., 2025; Zhang et al., 2025）为简化设计并实现零样本迁移，采用单一语义嵌入空间，这可能会低估局部结构的重要性，而这些结构对于从神经信号中忠实地重建图像至关重要。

跨模态对比学习 跨模态对比学习（CMCL）通过在温度缩放的 InfoNCE 目标下最大化匹配对之间的一致性，在共享嵌入空间中对齐异构输入（Oord 等人，2018；Wu 等人，2018；He 等人，2020）。采用余弦相似度和（通常为）对称损失的双编码器架构，已成为可扩展预训练和零样本迁移的默认方案（Radford 等人，2021；Jia 等人，2021；Zhai 等人，2022；Li 等人，2022）。基于这一方案，CLIP/ALIGN 等大规模视觉-语言系统在检索和分类基准上展现出强大的泛化能力，这一范式也从图像-文本扩展至音频-视觉（Arandjelovic 与 Zisserman，2017；Morgado 等人，2021；Wu 等人，2022）、视频-语言（Miech 等人，2019；2020；Xu 等人，2021；Bain 等人，2021；Luo 等人，2022）和 3D-语言（Xue 等人，2023；Zhang 等人，2022；Liu 等人，2023b）对齐。尽管取得了这些进展，CMCL 通常仍假设数据已准确配对。在网络规模的数据中，弱标注字幕、时间异步和域偏移会损害对齐质量，因此需要进行数据整理、字幕自举以及桥接/蒸馏策略（Jia 等人，2021；Miech 等人，2019）。近期分析（Liang 等人，2022；Wang 等人，2023）还揭示了共享空间中不同模态之间存在模态差异，这可能使细粒度对齐变得复杂。在神经解码中，脑-视觉配对天生稀缺且噪声较大（试次有限、信噪比低、试次间延迟存在差异），因此，若直接将脑信号对齐到单一的纯语义空间，可能会低估像素级结构，并加剧模态失配（Cichy 等人，2016a）。

图像扩散模型的适配器 适配器已成为参数高效模块，可在基本冻结基础权重的情况下，为预训练扩散模型扩展可控性和编辑能力，并为跨任务和跨模态应用提供统一方案（Wang 等，2025）。T2I-Adapter（Mou 等，2024）学习轻量级分支，将外部控制信号（例如边缘、深度和草图）与冻结文生图模型的内部特征对齐，从而实现准确且可组合的多条件控制。ControlNet（Zhang 等，2023）冻结预训练骨干网络，并添加零初始化侧网络，以注入空间条件，同时避免破坏原始先验。IP-Adapter（Ye 等，2023）将交叉注意力解耦，以便在文本之外整合图像提示；在保持扩散骨干网络冻结的同时，仅使用 2200 万个可训练参数即可实现强大的多模态条件控制。

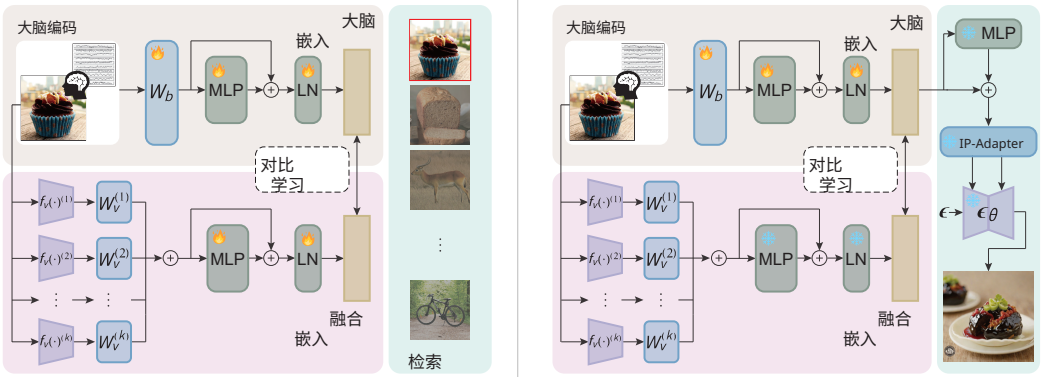
3 方法

3.1 问题陈述

视觉解码的目标是检索或重建与所记录脑信号相对应的视觉信息。我们将成对的脑信号和视觉图像表示为 $(x_v,\ x_b) \in \mathcal{D}$，其中 $x_v \in \mathbb{R}^{H\times W\times 3}$ 表示视觉刺激，H 和 W 分别表示图像的高度和宽度。$x_b \in \mathbb{R}^{C\times T}$ 表示在相同刺激下记录的脑信号，其中 C 对应电极通道的数量，T 表示时间采样点的长度。

3.2 将脑信号与层次化视觉表征对齐

直接将脑信号与视觉表征对齐，可能无法捕捉视觉信息固有的多尺度特性，从而限制对齐性能。尽管视觉模态中的高层语义特征对于类别识别和抽象理解至关重要，



（a）脑到图像检索。　　(b) 脑到图像重建。

图1：**学习流程**。左：检索目标，使用对称 InfoNCE 将大脑嵌入 z_b 与融合视觉嵌入 z_f （基于 K 预训练编码器的 HVF ）对齐；评估方式是在融合空间中进行最近邻检索。右：重建流程，采用冻结的预训练融合先验——HVF 加条件适配器（多层感知机投影器 + IP-Adapter，带有解耦交叉注意力）。我们将 z_b 与冻结的 z_f进行对比对齐，投影到 z_c，并将z_c 注入冻结的 SDXL UNet，以合成图像。视觉编码器和 UNet 均被冻结；对齐期间仅更新大脑侧。

低层特征则提供互补的结构信息和像素级细节，这些信息对于提升重建质量不可或缺。受这种视觉感知机制的启发（Blasdel & Lund, 1983; Tsumoto 等人, 1978），我们整合多个预训练视觉编码器，分别提取高层语义特征和低层像素特征，并通过对比学习目标将它们与脑信号嵌入对齐，从而构建统一的层次化视觉表征。

层次化视觉表征 如图1-(a)所示，我们设计了一种多头编码器结构，用于获取从高层视觉语义（例如物体、场景和关系）到低层视觉特征（例如颜色、纹理和布局）的层次化视觉表征。我们将 K 预训练编码器（默认使用 K=3 ）应用于图像 x_v，得到 $z_v^{(k)} = f_v^{(k)}(x_v)$ ，用于k=1, . . . , K。对于高层视觉语义，我们整合多个 CLIP 编码器，并从每个编码器中使用一个全局标记（即，基于 ViT 的模型使用 [CLS] 标记，基于 ResNet 的模型使用池化投影）。对于低层视觉特征，VAE 编码器输出形状为$[H/8, W/8, 4]$的潜变量，我们将其展平为长度为 $(H/8)(W/8) \times 4 = HW/16$的向量，同时保留局部结构和视觉细节。

我们使用后归一化残差层次化视觉融合器（HVF）融合特征。对于每个编码器，学习得到的线性映射 $W_{(k)\,v} \in \mathbb{R}^{d_k \times d}$将嵌入对齐到共享维度 d=1024:

$$\bar{z}_v = \sum_{k=1}^{K} z_v^{(k)} W_v^{(k)}, \tag{1}$$

对齐后的特征通过残差多层感知机（MLP）进行融合，得到

$$z_f = \mathrm{LayerNorm}\big(\bar{z}_v + \phi_v(\bar{z}_v)\big), \tag{2}$$

其中 ϕ_v 表示一个隐藏层大小为 $d_v = 1024$ 且采用 GELU 激活函数的两层多层感知机。

对比学习目标 对于脑模态，我们采用基于多层感知机的大脑投影（MBP）网络，将脑电信号投影到嵌入空间。我们首先将预处理后的信号展平为 $x_b' \in \mathbb{R}^{(C \cdot T)}$，然后通过学习得到的线性投影$W_b \in \mathbb{R}^{CT \times d}$将其映射到与视觉嵌入相同的维度。随后，我们沿用公式（2）中的相同架构，生成 d维嵌入，使其与 z_f 兼容，隐藏层大小为 d ，

$$\bar{z}_b = x_b' W_b, \qquad z_b = \mathrm{LayerNorm}\big(\bar{z}_b + \phi_b(\bar{z}_b)\big), \tag{3}$$

其中， ϕ_b 表示一个隐藏层大小为 $d_b = 1024$ 且采用 GELU 激活函数的两层多层感知机。

我们采用 CLIP式InfoNCE损失（Oord 等，2018）来对齐大脑与视觉嵌入。给定 N 对配对样本，我们使用可训练的温度参数 τ （初始化为 0.07）计算余弦相似度logits：

$$\hat{z}_b^{(i)} = \frac{z_b^{(i)}}{\|z_b^{(i)}\|_2}, \quad \hat{z}_f^{(i)} = \frac{z_f^{(i)}}{\|z_f^{(i)}\|_2}, \quad s_{ij} = \frac{\hat{z}_b^{(i)\top} \hat{z}_f^{(j)}}{\tau}, \tag{4}$$

其中， $\|\cdot\|_2$ 为 L2范数。学习目标定义为

$$\mathcal{L}_{\text{contrastive}} = -\frac{1}{2N}\left(\sum_{i=1}^{N} \log \frac{\exp(s_{ii})}{\sum_{j=1}^{N} \exp(s_{ij})} + \sum_{i=1}^{N} \log \frac{\exp(s_{ii})}{\sum_{j=1}^{N} \exp(s_{ji})}\right), \tag{5}$$

3.3 用于重建的预训练融合先验

虽然上述对比学习能够将脑信号与层次化视觉表征对齐，但直接将这些融合表征输入预训练扩散模型进行重建，往往会导致输出不稳定。核心问题在于缺少稳定的条件先验：脑信号驱动的特征尚未与生成模型预期的分布相匹配，从而导致引导信号带有噪声或未能对齐。为解决这一问题，我们引入融合先验，学习从融合视觉特征到扩散条件的稳健映射。

融合先验预训练 如图 1-(b) 所示，我们首先将来自 HVF 的融合视觉表征z_f 输入额外的投影器，以获得 z_c:

$$z_c = z_f + \phi_c(z_f), \tag{6}$$

其中 z_f， $z_c \in \mathbb{R}^d$ 且在公式 2 中， ϕ_v而 ϕ_c 表示一个隐藏层大小为 $d_c = 4096$且采用 GELU 激活函数的两层多层感知机。随后，IP-Adapter（Ye 等，2023）通过交叉注意力将 z_c 注入冻结的 SDXL（Podell 等，2023）UNet。给定时间步 t的含噪潜变量 x_t ，整个网络 δ 经过训练，以预测噪声 ϵ，并使用

$$\mathcal{L}_{\text{prior}} = \| \epsilon - \delta(x_t, t, z_c) \|_2^2, \tag{7}$$

其中 $\epsilon \sim \mathcal{N}(0, I)$ 是扩散目标， $\mathcal{L}_{\text{prior}}$ 是损失函数。

在大规模视觉数据上进行预训练时，UNet 骨干网络保持冻结，而 HVF 和投影器从头开始训练；IP-Adapter 则从预训练权重初始化[1] ，以加速收敛。文本提示留空，以确保模型学习从融合视觉特征到扩散条件的无文本映射。

大脑到融合空间的对齐 HVF 预训练完成后，我们将其冻结，仅使用与公式 (5) 相同的损失函数 $\mathcal{L}_{\text{contrastive}}$ 更新脑编码器，从而确保由大脑生成的嵌入被投影到稳定的预训练融合空间中。这可以防止表征漂移，并在输入扩散模型时实现稳健的重建。

重建完整流程 总的来说，训练分为两个阶段，推理分为一个阶段。(i) 先验预训练： 对于输入图像 x_v，提取 $\{z_v^{(k)}\}_{k=1}^{K}$，通过 HVF 和投影器将其融合并投影，得到 z_c；在空文本提示条件下，通过最小化式（7）中的$\mathcal{L}_{\text{prior}}$，联合训练 IP-Adapter、HVF 和投影器（UNet 冻结），从而获得稳定且不依赖文本的融合先验。(ii) 大脑—融合对齐： 冻结预训练的融合先验（HVF、投影器和 IP-Adapter）及 UNet，并仅更新大脑侧（即仅更新 MBP模块）；在配对数据 (x_b， x_v) 上使用式（5）中的对称 InfoNCE 损失，使 z_b 处于 z_f的融合空间中。(iii) 重建： 给定测试脑信号x_b，计算 $z_b = f_b(x_b)$，将其输入投影器以获得 z_c，并将 z_c 作为冻结的 IP-Adapter/UNet 的唯一条件；随后，标准扩散采样器（SDXL 使用 Euler祖先采样器（Karras 等人，2022））生成 $\hat{x}_v$，从而得到稳定且语义忠实的重建结果。

[1] https://huggingface.co/h94/IP-Adapter/resolve/main/sdxl_models/ ip-adapter_sdxl_vit-h.safetensors

4 实验

4.1 实验细节

我们在单张 NVIDIA 5090 32GB GPU 上训练对比阶段 25 个 epoch，全局批次大小为 1024。我们使用 AdamW，峰值学习率为 5×10^{-4}，采用余弦衰减调度，并从零开始进行 10 步预热。除非另有说明，检索使用固定的编码器组，包括 OpenAI CLIP RN50、LAION CLIP ViT-B/32（Schuhmann 等人，2022）和一个 SDXL VAE；每个骨干网络均采用其规范预处理。VAE 支持多种输入分辨率，默认分辨率为 128×128。对于生成，我们将 RN50 换为 LAION CLIP ViT-H/14，在视觉侧冻结预训练的 HVF，并且仅训练脑模态的 MBM模块。对比阶段在 THINGS-EEG 上训练（Grootswagers 等人，2019；Gifford 等人，2022b）



图 2：**困难样本检索对比**。我们的方法与 UBP 基线在困难案例集上检索到的前 5 张图像。

以及 THINGS-MEG（Hebart 等人，2023）。对于 THINGS-EEG（10 名参与者），训练划分包含 1654 个概念，每个概念有 10 张图像，每张图像重复 4 次；测试集划分包含 200 个概念，每个概念有 1 张图像，每张图像重复 80 次。我们遵循先前工作（Li 等，2024；Wu 等，2025），选取 17 个枕顶叶通道（O+P），并采用标准预处理方法（Song 等，2024）。对于 THINGS-MEG（4 名参与者，271 个通道），训练划分由$1854 \times 12 \times 1$（概念 × 图像 × 次重复）组成，测试集划分为 $200 \times 1 \times 12$。为提高信噪比（SNR），在两个数据集中，均按受试者对同一刺激的重复测量取平均值（训练集和测试集）。更多细节见附录 B。对于融合先验预训练，我们探索了多种先验配置。除非另有说明，训练使用两块 NVIDIA 5090 32 GB GPU，固定学习率为 1×10^{-4}，采用 SDXL-base 作为扩散骨干网络，并使用每块 GPU 可行的最大批量大小 12。每种配置均以 512×512 训练 100k 步，约两个训练轮次；每种先验配置耗时约 15 小时。预训练使用包含约 130 万张图像的 ImageNet-1k。推理时进行重建，我们使用 SDXL-Turbo 和四步采样器，以加快评估速度。

4.2 定量评估

我们评估两项任务：脑-视觉检索和脑-视觉重建。对于检索任务，我们报告在 THINGS-EEG 和 THINGS-MEG 上，采用受试者内和受试者间两种方案进行的 200 类零样本前1和前5准确率。对于重建任务，遵循先前的工作（Ozcelik & VanRullen，2023；Benchetrit 等人，2023；Li 等，2024），我们使用像素相关性和结构相似性指数衡量低层保真度，并采用这些工作中的其余语义和特征层面指标，包括

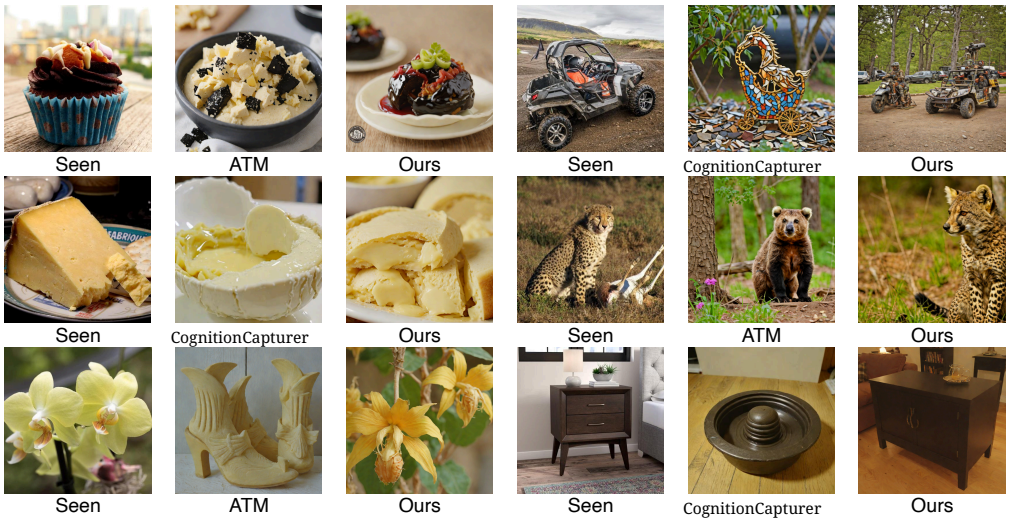


图 3：**脑到图像重建的定性比较**。每组三元组分别展示真实刺激（左）、基线（中）和我们的重建结果（右）。所有示例均使用受试者8的脑电图记录。

表 1：在 THINGS-**EEG** 和 THINGS-**MEG**上进行 200 类零样本检索的平均前1 / 前5准确率（%）。所有数值均为受试者平均值；“–”表示未报告。

方法	EEG				MEG			
	受试者内		受试者间		受试者内		受试者间	
	前1	前5	前1	前5	前1	前5	前1	前5
BraVL	5.8	17.5	1.8	7.0	–	–	–	–
NICE	16.1	43.6	6.2	21.4	12.8	36.0	–	–
NICE-SA	14.7	41.7	7.0	23.1	12.7	35.0	–	–
NICE-GA	15.6	42.8	5.9	21.6	14.3	42.3	–	–
MB2C	28.5	60.4	11.9	32.0	–	–	–	–
ATM	28.5	60.4	11.8	33.7	-	-	–	–
VE-SDN	37.2	69.9	–	–	–	–	–	–
CC-All	35.6	80.2	–	–	–	–	–	–
UBP	50.9	79.7	12.4	33.4	26.7	55.2	2.2	10.4
Ours	**75.7**	**94.6**	**20.0**	**44.1**	**33.7**	**60.5**	**5.4**	**15.2**

AlexNet(2/5)、Inception、CLIP 和 SwAV距离，其中数值越低越好。检索基线包括 BraVL（Du 等，2023）、NICE 及其空间变体（NICE-SA、NICE-GA）（Song 等，2024）、ATM（Li 等，2024）、VE-SDN（Chen 等，2024）、MB2C（Wei 等，2024）、UBP（Wu 等，2025）以及 CognitionCapturer（C.C.、All/Image/Depth/Text）（Zhang 等，2025）。对于重建，我们与 ATM（Li 等，2024）、CognitionCapturer（Zhang 等，2025）和 Brain Decoding（B.D.）（Benchetrit 等，2023）进行比较。当先前工作仅报告单被试结果时（例如 ATM 在受试者8上的结果），我们会在表格中注明。

与最强的先验工作（UBP）相比，我们的模型在所有协议下的200类零样本检索表现均稳定提升（前1/前5）：脑电图受试者内为75.7/94.6，对比50.9/79.7；脑电图受试者间为20.0/44.1，对比12.4/33.4；脑磁图受试者内为33.7/60.5，对比26.7/55.2；脑磁图受试者间为5.4/15.2，对比2.2/10.4（表1）。受试者间设置下的提升最大，表明跨参与者泛化能力更强。

表 2 总结了重建结果。在脑磁图上，我们的模型在低层相似性和语义对齐方面均达到或超过了先前工作，同时保持了有竞争力的 SwAV距离。在脑电图上，它改善了常见的受试者8案例，并在受试者平均表现上明显优于 ATM 和 C.C.。平均脑电图像素相关性从 C.C.（全部）的 0.150 提升至我们模型的 0.186，而结构相似性指数仍然相当；此外，在 AlexNet、Inception 和

表 2：脑电图和脑磁图重建质量的定量评估。

方法	数据集	低层		高层				
		像素相关性 ↑	结构相似性指数 ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	SwAV ↓
MEG	B.D.	0.076	0.336	0.736	0.826	0.671	0.767	**0.584**
	ATM	0.104	**0.340**	0.613	0.672	0.619	0.603	0.651
	Ours	**0.137**	0.292	**0.737**	**0.836**	**0.721**	**0.775**	0.600
EEG	C.C.（全部）	0.150	0.347	0.754	0.623	0.669	0.715	0.590
	C.C.（图像）	0.132	0.321	0.813	0.671	0.664	0.715	0.590
	C.C.（深度）	0.104	**0.370**	0.796	0.638	0.565	0.579	0.686
	C.C.（文本）	0.102	0.288	0.727	0.582	0.586	0.598	0.673
	Ours	**0.195**	0.336	**0.843**	**0.905**	**0.756**	**0.808**	**0.554**
脑电图（受试者-8）	ATM	0.160	0.345	0.776	0.866	0.734	0.786	0.582
	Ours	**0.227**	**0.361**	**0.878**	**0.924**	**0.796**	**0.826**	**0.531**

表 3：脑电检索消融实验：200 类零样本设置下的平均前1/前5准确率（%）；比较单个编码器、成对组合和三元组合。

设置	配置	受试者内		受试者间	
		前1	前5	前1	前5
单个模块	B32	52.2	83.3	13.3	33.9
	RN50	48.1	80.4	12.7	31.7
	VAE	44.3	75.2	10.2	23.9
两两组合	RN50 +B32	56.9	86.1	14.4	36.8
	RN50 +变分自编码器	65.8	90.4	17.4	37.3
	B32 + 变分自编码器	73.6	94.3	19.1	41.2
三重组合	RN50 +B32 + 变分自编码器	**75.7**	**94.6**	**20.0**	**44.1**

CLIP 上的语义相似性也有所提升，且 SwAV距离低于 C.C. 和 B.D.。综合来看，这些指标表明，我们的方法同时提升了保真度和语义一致性，而且这种改进并非仅限于单被试评估。

4.3 视觉比较

在图 2 中，我们展示了我们的方法和 UBP 基线在困难案例集上的前5个检索图像，结果表明我们的方法表现更好。我们还与先前的脑解码方法进行了定性比较。如图 3 所示，与 CognitionCapturer（Zhang 等，2025）和 ATM（Li 等，2024）相比，我们的方法重建出的图像具有更清晰的物体轮廓和更忠实的颜色分布。具体而言，我们的重建结果保留了细粒度的结构细节，同时捕捉到语义一致的属性，而这些属性往往是基线方法所缺失的。此外，整体感知质量与真实刺激更加吻合，展现了我们的框架在连接脑信号与视觉表征方面的有效性。

Ta表 4：针对脑到图像重建中不同融合先验效果的消融研究。

先验设置	像素相关性 ↑	结构相似性指数 ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	SwAV ↓
H14	0.174	0.327	0.825	0.872	0.733	0.773	0.574
H14 +B32	0.187	**0.340**	0.836	**0.908**	**0.783**	**0.814**	**0.547**
H14 +变分自编码器	0.173	0.312	0.789	0.838	0.672	0.721	0.611
H14 +B32 + 变分自编码器	**0.195**	0.336	**0.843**	0.905	0.756	0.808	0.554

4.4 消融研究

我们在表3中对视觉编码器组合在200类零样本脑电检索任务上的表现进行了消融研究。单个编码器提供了尚可的基线（例如，B32：受试者内前1/前5准确率为52.2/83.3；受试者间为13.3/33.9

表5：在 THINGS-EEG 上进行 200类零样本检索时，不同脑编码器骨干网络（行）和视觉编码器组合（列）的前1和前5准确率（%）；结果取所有 10名受试者的平均值，并使用相同的基于融合的视觉接口和超参数进行训练，仅改变脑骨干网络。完整 表示使用全部三个视觉编码器（RN50、B32 和 VAE）。

	RN50		B32		VAE		RN50+B32		RN50+变分自编码器		B32+变分自编码器		Full	
骨干网络	前1	前5	前1	前5	前1	前5	前1	前5	前1	前5	前1	前5		
ShallowNet	32.4	64.9	35.7	69.9	18.5	43.4	38.2	72.3	34.4	66.6	38.7	72.1	40.8	73.9
DeepNet	17.7	44.9	18.0	46.9	6.4	20.7	19.0	48.7	11.9	36.1	17.3	43.9	17.8	45.1
EEGNet	31.5	62.0	34.6	67.3	23.0	50.3	37.6	70.8	39.1	70.0	43.7	75.7	45.4	77.1
TSConv	38.7	72.0	42.2	74.5	25.8	55.6	45.5	78.1	46.4	78.1	52.5	83.0	54.5	84.1
ATM	41.5	74.1	42.9	76.0	34.4	63.9	46.8	78.2	52.2	81.4	56.2	85.4	57.9	86.5
BrainProjection	47.1	78.7	50.5	82.4	43.0	74.6	55.4	85.0	64.0	89.8	72.1	93.0	72.9	94.1
MBP（本文方法）	**48.1**	**80.5**	**51.2**	**83.0**	**44.1**	**75.5**	**56.2**	**86.2**	**65.2**	**90.7**	**73.7**	**94.1**	**75.7**	**94.6**



图 4：**融合先验消融实验**。每行展示了真实刺激，以及使用不同融合配置 H14、H14+B32、H14+VAE 和 H14+B32+VAE 生成的重建结果。所有示例均使用来自受试者 8 的脑电图记录。

），堆叠语义编码器（RN50+B32）带来了小幅提升（受试者内为56.9/86.1；受试者间为14.4/36.8）。加入 VAE 潜变量带来了最大幅度的提升：B32+VAE 达到73.6/94.3（受试者内）和19.1/41.2（受试者间），RN50+VAE 达到65.8/90.4（受试者内）和17.4/37.3（受试者间）。完整的 RN50+B32+VAE 融合在受试者内达到75.7/94.6，在受试者间达到20.0/44.1，在受试者内和受试者间划分上均较最佳双路设置带来小幅但稳定的提升。对于脑到图像的重建，表 4 比较了基于 H14、B32 和 VAE 特征构建的融合先验。在 H14 中加入 B32 提升了低层指标（像素相关性、结构相似性指数）和高层感知指标；完整的 H14+B32+VAE 先验进一步提升了像素相关性和 AlexNet 相似度，同时使其余指标保持在接近最佳双流设置的水平；图4展示了这些配置在定性表现上的代表性差异。

表 5 报告了在相同的 17 通道 O+P 设置和相同的基于融合的训练下的骨干网络比较。我们将 ShallowNet（Schirrmeister 等，2017）、DeepNet（Schirrmeister 等，2017）、EEGNet（Lawhern 等，2018）、TSConv（Song 等，2024）、ATM（Li 等，2024）、BrainProjection（Wu 等，2025）以及我们的 MBP 接入融合视觉接口（RN50、B32、变分自编码器及其组合），并固定视觉编码器、对齐目标和损失函数。所有先前的骨干网络

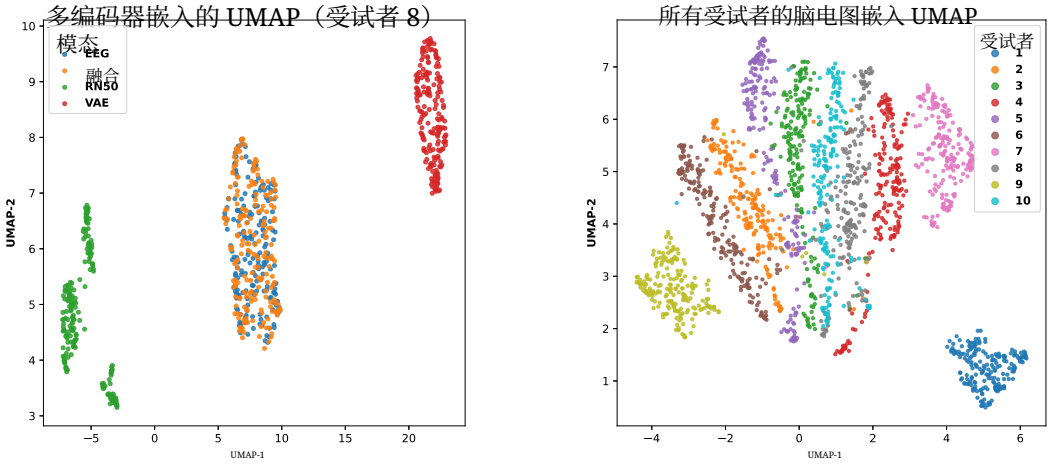


图 5：THINGS-EEG 数据集测试集划分上学习得到的嵌入的 UMAP 可视化。左：多编码器视觉嵌入（RN50、展平的变分自编码器）和融合令牌与受试者 8 的脑电图嵌入一同投影。右：全部 10 名受试者的脑电图嵌入。

其前1/前5检索准确率均高于原始单编码器报告中的结果，而 MBP 在每种视觉编码器组合中都取得了最高准确率；在 Full RN50+B32+变分自编码器配置下，其准确率达到 75.7%/94.6%，相比之下 BrainProjection 为 72.9%/94.1%，表明同一融合接口可以在不同脑电图骨干网络之间即插即用。

4.5 Q学习得到的嵌入的定性分析 INGS

图5展示了学习得到的嵌入，这些嵌入是在训练划分上训练后，于 THINGS-EEG 测试集划分上评估的。左图中，受试者 8 的脑电图嵌入非常接近融合视觉标记，且远离 RN50 和变分自编码器嵌入，这表明脑电图与融合表征对齐，而非与任何单一编码器对齐。右图中，10名受试者的脑电图嵌入形成了明显分离的簇，表明存在受试者特有的语义分布，这与脑电图中通常观察到的显著受试者间变异相符。附录中的图7显示脑电图嵌入与图像嵌入之间存在清晰的对角线，证实学习得到的融合表征可在测试集上支持可靠的200类零样本检索。

5 结论

本文研究了利用分层视觉嵌入学习脑表征，以实现脑到图像解码。我们提出了一种基于融合的脑视觉接口，将脑信号对齐到一个由互补的语义编码器和像素级编码器构建的单一词元，并将其输入冻结的生成先验。在 THINGS-EEG 和 THINGS-MEG 上的实验表明，该接口在受试者内和受试者间设置下均能实现出色的 200类零样本检索，并在受试者内解码中生成高质量的重建结果。消融研究表明，与单编码器基线和仅采用语义信息的基线相比，融合 CLIP 和变分自编码器特征能够提升脑到图像解码性能。将各种脑电图骨干网络接入同一接口，无需重新训练视觉侧即可稳定提升检索效果，这凸显了分层视觉嵌入是一种即插即用的途径，可为脑到视觉解码构建稳健的脑表征。

局限性与未来工作 我们的实验仅限于 THINGS-EEG 和 THINGS-MEG；将所提出的接口扩展至其他数据集、模态和任务，是未来工作的重要方向。当前的融合设计和视觉堆栈采用的是单一的人工选定配置，探索更强大的视觉骨干网络和其他编码器组，或许能进一步提升解码性能和效率。

致谢

本研究得到国家自然科学基金委员会资助，项目编号为 62502411。

参考文献

Pranav Aggarwal、Hareesh Ravi、Naveen Marri、Sachin Kelkar、Fengbin Chen、Vinh Khuc、Midhun Harikumar、Ritiz Tambi、Sudharshan Reddy Kakumanu、Purvak Lapsiya 等。基于扩散先验的受控条件文本到图像生成。*arXiv*预印本 *arXiv:2302.11710*，2023。

Relja Arandjelovic 和 Andrew Zisserman。看、听与学习。载于*IEEE* 国际计算机视觉会议论文集，第 609–617 页，2017年。

Yunpeng Bai、Xintao Wang、Yan-pei Cao、Yixiao Ge、Chun Yuan 和 Ying Shan。Dreamdiffusion：从脑电信号生成高质量图像。*arXiv* 预印本 *arXiv:2306.16934*，2023。

Max Bain、Arsha Nagrani、G¨ul Varol 和 Andrew Zisserman。《Frozen in time：用于端到端检索的视频和图像联合编码器》。载于*IEEE/CVF*国际计算机视觉大会论文集，第1728–1738页，2021年。

Hangbo Bao、Li Dong、Songhao Piao 和 Furu Wei。Beit：图像Transformer的BERT预训练。*arXiv*预印本 *arXiv:2106.08254*，2021年。

Yohann Benchetrit、Hubert Banville 和 Jean-R´emi King。脑解码：迈向视觉感知的实时重建。*arXiv*预印本 *arXiv:2310.19812*，2023。

Gary G Blasdel 和 Jennifer S Lund。猕猴纹状皮层中传入轴突的终止。神经科学杂志，3(7):1389–1413，1983年。

Mathilde Caron、Hugo Touvron、Ishan Misra、Herv´e J´egou、Julien Mairal、Piotr Bojanowski 和 Armand Joulin。自监督视觉Transformer中的新兴特性。载于*IEEE/CVF*国际计算机视觉大会论文集，第9650–9660页，2021年。

Hongzhou Chen、Lianghua He、Yihang Liu 和 Longzhen Yang。通过改进视觉-脑电图语义一致性实现视觉神经解码。*arXiv*预印本 *arXiv:2408.06788*，2024。

Radoslaw Martin Cichy、Aditya Khosla、Dimitrios Pantazis、Antonio Torralba 和 Aude Oliva。将深度神经网络与人类视觉物体识别的时空皮层动态进行比较，揭示了层级对应关系。《科学报告》，6(1):27755，2016a。

Radoslaw Martin Cichy、Dimitrios Pantazis 和 Aude Oliva。基于相似性的脑磁图与功能性磁共振成像融合揭示了视觉物体识别过程中人类大脑皮层的时空动态。《大脑皮层》，26(8):3563–3579，2016b。

Fernando Lopes da Silva。脑电图和脑磁图：与神经科学的相关性。《神经元》，80(5):1112–1128，2013。

Alexey Dosovitskiy、LucasBeyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly 等。一幅图像胜过 16x16 个词：用于大规模图像识别的 Transformer。*arXiv* 预印本 *arXiv:2010.11929*，2020年。

Changde Du、Kaicheng Fu、Jinpeng Li 和 Huiguang He。通过脑-视觉-语言特征的多模态学习解码视觉神经表征。《IEEE模式分析与机器智能汇刊》，45(9):10760–10777，2023。

Patrick Esser、Sumith Kulal、Andreas Blattmann、Rahim Entezari、Jonas M¨uller、Harry Saini、Yam Levi、Dominik Lorenz、Axel Sauer、Frederic Boesel 等。扩展整流流Transformer以实现高分辨率图像合成。发表于第四十一届国际机器学习会议，2024年。

Honghao Fu、Hao Wang、Jing Jih Chin 和 Zhiqi Shen。Brainvis：通过图像重建探索大脑与视觉信号之间的桥梁。发表于*ICASSP 2025-2025 IEEE*国际声学、语音与信号处理会议 *(ICASSP)*，第1–5页。IEEE，2025年。

Alessandro T. Gifford、Kshitij Dwivedi、Gemma Roig 和 Radoslaw M. Cichy。用于建模人类视觉物体识别的大型丰富脑电图数据集。*NeuroImage*，264:119754，2022a。ISSN 1053-8119。doi: https://doi.org/10.1016/j.neuroimage.2022.119754。URLhttps: //www.sciencedirect.com/science/article/pii/S1053811922008758。

Alessandro T Gifford、Kshitij Dwivedi、Gemma Roig 和 Radoslaw M Cichy。用于建模人类视觉物体识别的大型丰富脑电图数据集。*NeuroImage*，264:119754，2022b。

Tijl Grootswagers、AmandaK Robinson 和 Thomas A Carlson。快速连续视觉加工流中视觉物体的表征动态。*NeuroImage*，188:668–679，2019年。

何恺明、Haoqi Fan、Yuxin Wu、Saining Xie 和 Ross Girshick。用于无监督视觉表征学习的动量对比。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 9729–9738 页，2020年。

何恺明、Xinlei Chen、Saining Xie、Yanghao Li、Piotr Doll´ar 和 Ross Girshick。掩码自编码器是可扩展的视觉学习器。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 16000–16009 页，2022。

Martin N Hebart、Oliver Contier、Lina Teichmann、Adam H Rockter、Charles Y Zheng、Alexis Kid- der、Anna Corriveau、Maryam Vaziri-Pashkam 和 Chris I Baker。THINGS-data：用于研究人脑与行为中物体表征的大规模数据集多模态合集。*eLife*，12:e82580，2023。

Irina Higgins、Loic Matthey、Arka Pal、Christopher Burgess、Xavier Glorot、Matthew Botvinick、Shakir Mohamed 和 Alexander Lerchner。β-VAE：利用受约束的变分框架学习基本视觉概念。发表于国际学习表征会议，2017年。

Chao Jia、Yinfei Yang、Ye Xia、Yi-Ting Chen、Zarana Parekh、Hieu Pham、Quoc Le、Yun-Hsuan Sung、Zhen Li 和 Tom Duerig。利用噪声文本监督扩展视觉与视觉-语言表征学习。发表于国际机器学习会议，第 4904–4916 页。PMLR，2021年。

Tero Karras、Miika Aittala、Timo Aila 和 Samuli Laine。阐明基于扩散的生成模型的设计空间。神经信息处理系统进展，35：26565–26577，2022。

Kendrick NKay、Thomas Naselaris、Ryan J Prenger 和 Jack L Gallant。从人脑活动中识别自然图像。《自然》，452(7185):352–355，2008。

Diederik PKingma 和 Max Welling。自编码变分贝叶斯。*arXiv* 预印本 *arXiv:1312.6114*，2013。

Vernon J Lawhern、Amelia J Solon、Nicholas R Waytowich、Stephen M Gordon、Chou P Hung 和 Brent J Lance。EEGNet：一种用于基于脑电的脑机接口的紧凑型卷积神经网络。《神经工程杂志》，15(5):056013，2018。

Dongyang Li、Chen Wei、Shiying Li、Jiachen Zou 和 Quanying Liu。利用带引导扩散的脑电图嵌入进行视觉解码与重建。在第*38*届国际神经信息处理系统会议论文集中，第102822–102864页，2024。

Haoyu Li、Hao Wu 和 Badong Chen。Neuraldiffuser：用于功能磁共振成像视觉重建的神经科学启发的扩散引导。《*IEEE*图像处理汇刊》，2025。

Junnan Li、Dongxu Li、Caiming Xiong 和 Steven Hoi。BLIP：通过自举式语言-图像预训练，实现统一的视觉-语言理解与生成。发表于国际机器学习会议，第 12888–12900 页。PMLR，2022。

Victor Weixin Liang、Yuhui Zhang、Yongchan Kwon、Serena Yeung 和 James Y Zou。注意模态差异：理解多模态对比表征学习中的模态差异。发表于神经信息处理系统进展，第 35 卷：17612–17625，2022。

Fang Liu、Pei Yang、Yezhi Shu、Niqi Liu、Jenny Sheng、Junwen Luo、Xiaoan Wang 和 Yong-Jin Liu。通过整合深度特征聚合与迁移学习，基于少通道脑电信号进行情绪识别。*IEEE*情感计算汇刊，15(3):1315–1330，2023a。

Minghua Liu、Ruoxi Shi、Kaiming Kuang、Yinhao Zhu、Xuanlin Li、Shizhong Han、Hong Cai、Fatih Porikli 和 Hao Su。OpenShape：扩展三维形状表征以实现开放世界理解。神经信息处理系统进展，36:44860–44879，2023b。

Yulong Liu、Yongqiang Ma、Wei Zhou、Guibo Zhu 和 Nanning Zheng。BrainCLIP：通过 CLIP 弥合脑与视觉语言表征，以实现通用自然视觉刺激解码。*arXiv* 预印本 *arXiv:2302.12971*，2023c。

Nikos KLogothetis、Jon Pauls、Mark Augath、Torsten Trinath 和 Axel Oeltermann。对功能性磁共振成像信号基础的神经生理学研究。《自然》，412(6843):150–157，2001年。

Huaishao Luo、Lei Ji、Ming Zhong、Yang Chen、Wen Lei、Nan Duan 和 Tianrui Li。Clip4clip：一项关于利用 CLIP 进行端到端视频片段检索与字幕生成的实证研究。《神经计算》，508：293–304，2022。

WilliamH Merigan和 JH Maunsell。灵长类视觉通路的并行程度如何？《神经科学年度评论》，1993年。

Antoine Miech、Dimitri Zhukov、Jean-Baptiste Alayrac、Makarand Tapaswi、Ivan Laptev 和 Josef Sivic。Howto100m：通过观看一亿个配有旁白的视频片段学习文本-视频嵌入。载于*IEEE/CVF*国际计算机视觉大会论文集，第2630–2640页，2019年。

Antoine Miech、Jean-Baptiste Alayrac、Lucas Smaira、Ivan Laptev、Josef Sivic 和 Andrew Zisserman。从未经筛选的教学视频中端到端学习视觉表征。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 9879–9889 页，2020年。

Pedro Morgado、Nuno Vasconcelos 和 Ishan Misra。通过跨模态一致性进行音视实例判别。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 12475–12486 页，2021年。Chong Mou、Xintao Wang、Liangbin Xie、Yanze Wu、Jian Zhang、Zhongang Qi 和 Ying Shan。T2i-adapter：学习适配器，挖掘文生图扩散模型更强的可控能力。载于*AAAI*人工智能会议论文集，第 38 卷，第 4296–4304 页，2024。Aaron van den Oord、Yazhe Li 和 Oriol Vinyals。基于对比预测编码的表征学习。*arXiv*预印本 *arXiv:1807.03748*，2018。Maxime Oquab、Timoth´ee Darcet、Th´eo Moutakanni、Huy Vo、Marc Szafraniec、Vasil Khalidov、Pierre Fernandez、Daniel Haziza、Francisco Massa、Alaaeldin El-Nouby 等。DINOv2：无监督学习稳健视觉特征。*arXiv*预印本 *arXiv:2304.07193*，2023。Furkan Ozcelik 和 Rufin VanRullen。利用生成式潜在扩散从功能性磁共振成像信号重建自然场景。《科学报告》，13(1):15666，2023。Dustin Podell、Zion English、Kyle Lacey、Andreas Blattmann、Tim Dockhorn、Jonas M¨uller、Joe Penna 和 Robin Rombach。SDXL：改进潜在扩散模型，实现高分辨率图像合成。*arXiv*预印本 *arXiv:2307.01952*，2023。Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。通过自然语言监督学习可迁移视觉模型。载于国际机器学习会议，第 8748–8763 页。PMLR，2021。

Ali Razavi、Aaron Van den Oord 和 Oriol Vinyals。利用 vq-vae-2 生成多样且高保真度的图像。神经信息处理系统进展，第 32 卷，2019年。

Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Bj¨orn Ommer。基于潜在扩散模型的高分辨率图像合成。载于 *IEEE/CVF* 计算机视觉与模式识别会议论文集，第 10684–10695 页，2022。

Robin Tibor Schirrmeister、Jost Tobias Springenberg、Lukas Dominique Josef Fiederer、Martin Glasstetter、Katharina Eggensperger、Michael Tangermann、Frank Hutter、Wolfram Burgard 和 Tonio Ball。使用卷积神经网络进行深度学习、脑电图解码与可视化。人脑图谱，38(11)：5391–5420，2017年。

Christoph Schuhmann、Romain Beaumont、Richard Vencu、Cade Gordon、Ross Wightman、Mehdi Cherti、Theo Coombes、Aarush Katta、Clayton Mullis、Mitchell Wortsman 等。LAION-5B：用于训练新一代图文模型的开放大规模数据集。神经信息处理系统进展，35：25278–25294，2022。

Yonghao Song、Bingchuan Liu、Xiang Li、Nanlin Shi、Yijun Wang 和 Xiaorong Gao。利用脑电图解码自然图像以进行目标识别。载于第十二届国际学习表征会议，2024。URL https://openreview.net/forum?id=dhLIno8FmH。

Yu Takagi 和 Shinji Nishimoto。利用人脑活动和潜在扩散模型进行高分辨率图像重建。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 14453–14463 页，2023a。

Yu Takagi 和 Shinji Nishimoto。通过多个解码输入，利用潜在扩散模型改进基于人脑活动的视觉图像重建。*arXiv* 预印本 *arXiv:2306.11536*，2023b。

Yonglong Tian、Lijie Fan、Kaifeng Chen、Dina Katabi、Dilip Krishnan 和 Phillip Isola。向模型学习视觉，其效果可与从数据中学习视觉相媲美。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 15887–15898 页，2024。

T Tsumoto、OD Creutzfeldt和CR Legendy。猫视觉皮层至外侧膝状体的皮质传出系统功能组织：附关于膝状体-皮层单突触连接的附录。实验脑研究，32(3):345–364，1978年。

Aaron Van Den Oord、Oriol Vinyals等。神经离散表征学习。神经信息处理系统进展，30，2017年。

Jia Wang、Jie Hu、Xiaoqi Ma、Hanghang Ma、Xiaoming Wei和Enhua Wu。基于扩散模型的图像编辑：综述。*arXiv*预印本 *arXiv:2504.13226*，2025。

Shizun Wang、Songhua Liu、Zhenxiong Tan和Xinchao Wang。Mindbridge：一种跨被试脑解码框架。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第11333–11342页，2024。

Zehan Wang、Yang Zhao、Haifeng Huang、Jiageng Liu、Aoxiong Yin、Li Tang、Linjun Li、Yongqi Wang、Ziang Zhang 和 Zhou Zhao。连接多模态对比表征。神经信息处理系统进展，36：22099–22114，2023。

Yayun Wei、Lei Cao、Hao Li 和 Yilin Dong。MB2C：用于学习鲁棒视觉神经表征的多模态双向循环一致性。载于第*32*届*ACM*国际多媒体会议论文集，第8992–9000页，2024。

Haitao Wu、Qing Li、Changqing Zhang、Zhen He 和 Xiaomin Ying。利用不确定性感知模糊先验弥合视觉-大脑差距。载于计算机视觉与模式识别会议论文集，第2246–2257页，2025。

Ho-Hsiang Wu、Prem Seetharaman、Kundan Kumar 和 Juan Pablo Bello。Wav2clip：从 clip 中学习鲁棒的音频表征。载于*ICASSP 2022-2022 IEEE*国际声学、语音与信号处理会议 *(ICASSP)*，第 4563–4567 页。IEEE，2022。

Zhirong Wu、Yuanjun Xiong、Stella X Yu 和 Dahua Lin。通过非参数实例判别进行无监督特征学习。载于*IEEE*计算机视觉与模式识别会议论文集，第 3733–3742 页，2018。

Zhenda Xie、Zheng Zhang、Yue Cao、Yutong Lin、Jianmin Bao、Zhuliang Yao、Qi Dai 和 Han Hu。SimMIM：用于掩码图像建模的简单框架。载于 *IEEE/CVF*计算机视觉与模式识别会议论文集，第 9653–9663 页，2022。

Hu Xu、Gargi Ghosh、Po-Yao Huang、Dmytro Okhonko、Armen Aghajanyan、Florian Metze、Luke Zettlemoyer 和 Christoph Feichtenhofer。VideoCLIP：用于零样本视频-文本理解的对比式预训练。*arXiv* 预印本 *arXiv:2109.14084*，2021年。

Xiran Xu、Bo Wang、Boda Xiao、Yadong Niu、Yiwen Wang、Xihong Wu 和 Jing Chen。警惕由脑电图信号中的时间自相关导致的过高解码性能。*arXiv* 预印本 *arXiv:2405.17024*，2024。

Le Xue、Mingfei Gao、Chen Xing、Roberto Mart´ın-Mart´ın、Jiajun Wu、Caiming Xiong、Ran Xu、Juan Carlos Niebles 和 Silvio Savarese。ULIP：为三维理解学习语言、图像和点云的统一表征。发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 1179–1189 页，2023 年。

Daniel LK Yamins、Ha Hong、Charles F Cadieu、Ethan A Solomon、Darren Seibert 和 James J DiCarlo。经过性能优化的层次模型可预测高级视觉皮层中的神经反应。美国国家科学院院刊，111(23):8619–8624，2014 年。

Hu Ye、Jun Zhang、Sibo Liu、Xiao Han 和 Wei Yang。IP-Adapter：兼容文本的图像提示适配器，适用于文生图扩散模型。*arXiv* 预印本 *arXiv:2308.06721*，2023 年。

Xiaohua Zhai、Xiao Wang、Basil Mustafa、Andreas Steiner、Daniel Keysers、Alexander Kolesnikov 和 Lucas Beyer。LiT：通过锁定图像文本调优实现零样本迁移。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 18123–18133 页，2022。

Kaifan Zhang、Lihuo He、Xin Jiang、Wen Lu、Di Wang 和 Xinbo Gao。CognitionCapturer：利用多模态信息从人类脑电信号中解码视觉刺激。载于*AAAI*人工智能会议论文集，第 39 卷，第 14486–14493 页，2025。

Lvmin Zhang、Anyi Rao 和 Maneesh Agrawala。为文生图扩散模型添加条件控制。载于*IEEE/CVF*国际计算机视觉大会论文集，第 3836–3847 页，2023。

Renrui Zhang、Ziyu Guo、Wei Zhang、Kunchang Li、Xupeng Miao、Bin Cui、Yu Qiao、Peng Gao 和 Hongsheng Li。Pointclip：基于 CLIP 的点云理解。发表于*IEEE/CVF*计算机视觉与模式识别会议，第 8552–8562 页，2022。

A 大语言模型使用声明

我们使用大语言模型进行语法检查和语言润色，以提高可读性。

B 数据集详情

THINGS-EEG THINGS-EEG（Gifford 等人，2022a）是一个大型数据集，包含 10 名参与者的脑电图（EEG）记录。信号由一顶按照国际 10–10 系统排列的 64 通道 EASYCAP 采集。训练划分涵盖 1,654 个物体概念，每个概念由 10 张图像表示；每张图像向每位参与者呈现四次（$1{,}654 \times 10 \times 4$）。测试集划分涵盖 200 个概念，每个概念包含一张图像，重复呈现 80 次（$200 \times 1 \times 80$）。预处理遵循 Song 等人（2024）和 Wu 等人（2025）的方法：将原始脑电图滤波至 0.1–100 Hz，得到采样率为 1,000 Hz 的 63 个通道；将试次截取为刺激后 0–1,000 ms 的片段，并使用此前 200 ms 的平均值进行基线校正。随后将数据降采样至 250 Hz，并保留枕叶和顶叶区域中与视觉处理相关的 17 个后部通道。为提高信噪比，对重复试次取平均值，从而为每位参与者生成 16,540 个训练样本和 200 个测试样本。

THINGS-MEG THINGS-MEG（Hebart 等人，2023）是一个大规模数据集，包含 4 名参与者的脑磁图（MEG）记录。信号通过 271 个通道采集。每个试次呈现一张图像 500 ms，随后显示空白屏幕 1000 ± 200ms。训练划分涵盖 1,854 个物体概念，每个概念由 12 张图像表示；每张图像向每位参与者呈现一次（$1{,}854 \times 12 \times 1$）。测试集划分包含 200 个概念，每个概念对应一张图像，并重复呈现 12 次（$200 \times 1 \times 12$）。为构建零样本任务，将 200 个测试概念从训练集中剔除。预处理遵循 Song 等人（2024）；Wu 等人（2025）的方法：原始脑磁图信号经过 0.1–100 Hz 滤波；试次截取刺激后 0–1,000 ms 的时间段，并进行基线校正。随后将数据降采样至 200 Hz。为提高信噪比，对重复试次取平均，使每位参与者得到 19,848 个训练样本和 200 个测试样本。

伦理考量 我们的实验完全依赖于公开发布的 THINGS- EEG/MEG 数据集（Gifford 等人，2022a；Hebart 等人, 2023）。这些数据集是在获得知情同意并接受机构监督的情况下收集的，我们仅使用匿名化记录。尽管脑到图像的解码最终可能使闭锁综合征患者或言语障碍患者等群体受益，但也会带来明显风险，包括未经授权地推断心理状态以及隐私侵犯。我们将本方法视为针对基准数据集的一项技术贡献，而非秘密监控工具，并强调任何现实应用都必须获得明确同意、采取稳健的数据保护措施，并遵守临床与法律法规。

C 结果详情

THINGS-EEG 和 THINGS-MEG 上的被试内检索 我们报告了 THINGS-EEG 和 THINGS-MEG 中每位受试者在 200 类零样本设置下的 Top-1/Top-5 准确率。对于每位受试者，我们评估了单个编码器（RN50、B32、变分自编码器）、两两堆叠（RN50+B32、RN50+变分自编码器、B32+变分自编码器）以及三元堆叠（RN50+B32+变分自编码器），其中变分自编码器的输入固定为 128×128。

表6：THINGS-EEG 上 200类零样本检索的前1和前5准确率（%）。

方法	Sub1		Sub2		Sub3		Sub4		Sub5		Sub6		Sub7		Sub8		Sub9		Sub10		Avg	
	前1	前5	前1	前5	前1	前5	前1	前5	前1	前5	前1	前5	前1	前5	前1	前5	前1	前5	前1	前5	前1	前5
BraVL	6.1	17.9	4.9	14.9	5.6	17.4	5.0	15.1	4.0	13.4	6.0	18.2	6.5	20.4	8.8	23.7	4.3	14.0	7.0	19.7	5.8	17.5
NICE	13.2	39.5	13.5	40.3	14.5	42.7	20.6	52.7	10.1	31.5	16.5	44.0	17.0	42.1	22.9	56.1	15.4	41.6	17.4	45.8	16.1	43.6
NICE-SA	13.3	40.2	12.1	36.1	15.3	39.6	15.9	49.0	9.8	34.4	14.2	42.4	17.9	43.6	18.2	50.2	14.4	38.7	16.0	42.8	14.7	41.7
NICE-GA	15.2	40.1	13.9	40.1	14.7	42.7	17.6	48.9	9.0	29.7	16.4	44.4	14.9	43.1	20.3	52.1	14.1	39.7	19.6	46.7	15.6	42.8
MB2C	23.7	56.3	22.7	50.5	26.3	60.2	34.8	67.0	21.3	53.0	31.0	62.3	25.0	54.8	39.0	69.3	27.5	59.3	33.2	70.8	28.5	60.4
ATM-S	25.6	60.4	22.0	54.5	25.0	62.4	31.4	60.9	12.9	43.0	21.3	51.1	30.5	61.5	38.8	72.0	34.4	51.5	29.1	63.5	28.5	60.4
VE-SDN	32.6	63.7	34.4	69.9	38.7	73.5	39.8	72.0	29.4	58.6	34.5	68.8	34.5	68.3	49.3	79.8	39.0	69.6	39.8	75.3	37.2	69.9
CognitionCapturer-All	31.4	79.7	31.4	77.8	38.2	85.7	40.4	85.8	24.4	66.3	34.8	78.8	34.7	81.0	48.1	88.6	31.4	79.4	35.6	79.3	35.6	80.2
UBP	41.2	70.5	51.2	80.9	51.2	82.0	51.1	76.9	42.2	72.8	57.5	83.5	49.0	79.9	58.6	85.8	45.1	76.2	61.5	88.2	50.9	79.7
Ours	64.3	88.8	76.3	95.3	74.0	95.0	67	91.8	68.0	91.5	81.5	96.3	76.8	96.8	84.8	98.5	76.8	95.8	87.3	99.3	75.7	94.6

逐被试重建指标 我们进一步在表10、表11、表12和表13中报告了各受试者的重建指标。对于每位受试者，我们计算低级指标（像素相关性，

表 7：不同配置下 THINGS-EEG 上 200类零样本检索的前1和前5准确率（%）。

配置	Sub1		Sub2		Sub3		Sub4		Sub5		Sub6		Sub7		Sub8		Sub9		Sub10		Avg	
	前1	前5	前1	前5	前1	前5	前1	前5	前1	前5	前1	前5	前1	前5	前1	前5	前1	前5	前1	前5	前1	前5
B32	39.3	75.3	48.8	79.3	53.3	84.5	54.8	87.0	42.8	75.0	57.8	84.3	47.0	81.0	62.3	89.3	44.3	80.0	61.8	94.3	51.2	83.0
RN50	40.0	69.5	48.5	79.5	48.5	85.0	45.8	82.0	41.5	74.0	55.8	83.0	48.5	77.5	55.5	88.0	41.8	77.0	55.3	89.5	48.1	80.5
VAE	38.8	71.5	41.3	72.5	43.3	75.3	33.5	65.3	39.8	70.5	50.5	81.8	44.3	75.0	55.0	86.3	42.8	73.3	52.0	83.5	44.1	75.5
RN50+B32	47.3	79.0	55.8	81.0	56.3	87.5	59.8	88.8	46.8	81.0	63.0	87.5	53.0	85.0	65.0	91.3	50.0	86.5	68.3	94.5	56.5	86.2
RN50+变分自编码器	60.8	86.0	62.8	92.0	59.8	91.0	53.3	87.0	58.0	84.0	73.0	94.0	62.5	88.8	77.0	97.3	68.3	90.8	77.0	96.5	65.2	90.7
B32+变分自编码器	63.0	88.3	70.3	94.0	73.5	94.3	64.3	92.3	70.5	91.0	78.0	96.0	73.3	93.3	84.8	97.5	75.0	96.0	84.8	98.8	73.7	94.1
RN50+B32+变分自编码器	64.3	88.8	76.3	95.3	74.0	95.0	67	91.8	68.0	91.5	81.5	96.3	76.8	96.8	84.8	98.5	76.8	95.8	87.3	99.3	75.7	94.6

表8：THINGS-MEG上200类零样本检索的前1和前5准确率（%）

方法	Sub1		Sub2		Sub3		Sub4		Avg	
	前1	前5	前1	前5	前1	前5	前1	前5		
NICE	9.6	27.8	18.5	47.8	14.2	41.6	9.0	26.6	12.8	36.0
NICE-SA	9.8	27.8	18.6	46.4	10.5	38.4	11.7	27.2	12.7	35.0
NICE-GA	8.7	30.5	21.8	56.6	16.5	49.7	10.3	32.3	14.3	42.3
UBP	15.0	38.0	46.0	80.5	27.3	59.0	18.5	43.5	26.7	55.2
Ours	14.0	31.8	63.8	91.8	41.0	78.3	17.0	41.0	33.9	60.7

结构相似性指数）和高层感知相似度（AlexNet(2/5)、Inception、CLIP），并使用 SwAV↓ 作为多样性/一致性代理指标。结果分别展示了单一目标（仅 H14）、语义对（H14+B32、H14+变分自编码器）和完整的多尺度堆栈（H14+B32+变分自编码器）。最后一行给出了各受试者的均值。

不同受试者的重建如图6所示，对于相同的视觉刺激，我们根据不同受试者记录的脑电图重建图像。



图6：**跨受试者脑电图重建**。

表 9：不同配置下，THINGS-MEG 上 200类零样本检索的前1和前5准确率（%）。

配置	Sub1		Sub2		Sub3		Sub4		Avg	
	前1	前5	前1	前5	前1	前5	前1	前5	前1	前5
B32	9.8	31.5	52.8	81.3	31.8	67.0	19.5	47.5	28.4	56.8
RN50	12.0	37.5	50.3	83.0	29.8	65.8	19.0	44.5	27.8	57.7
VAE	2.9	13.2	32.5	65.3	12.4	39.1	5.6	19.1	13.4	34.2
RN50+B32	9.8	31.5	52.0	83.0	32.5	67.8	18.8	47.8	28.3	57.5
RN50+变分自编码器	9.0	22.8	48.0	85.3	26.5	61.3	11.5	30.3	23.8	49.9
B32+变分自编码器	12.0	33.5	64.5	91.3	39.0	76.8	17.3	43.8	33.2	61.3
RN50+B32+变分自编码器	14.0	31.8	63.8	91.8	41.0	78.3	17.0	41.0	33.9	60.7

表10：使用H14设置时各受试者的重建指标（↑越高越好，↓越低越好）。

受试者	低层		高层				
	像素相关性↑	结构相似性指数↑	AlexNet(2)↑	AlexNet(5)↑	Inception↑	CLIP↑	SwAV↓
1	0.179	0.305	0.828	0.870	0.719	0.732	0.588
2	0.174	0.331	0.826	0.868	0.712	0.769	0.588
3	0.177	0.317	0.832	0.872	0.703	0.802	0.574
4	0.167	0.326	0.803	0.863	0.752	0.778	0.573
5	0.163	0.315	0.804	0.846	0.676	0.745	0.593
6	0.181	0.316	0.838	0.874	0.715	0.763	0.588
7	0.155	0.328	0.811	0.874	0.736	0.775	0.569
8	0.193	0.349	0.852	0.906	0.781	0.795	0.550
9	0.163	0.330	0.820	0.872	0.765	0.762	0.561
10	0.192	0.350	0.837	0.870	0.773	0.810	0.556
Ave	0.174	0.327	0.825	0.871	0.733	0.773	0.574

表 11：使用 H14+B32 设置时各受试者的重建指标（ ↑ 越高越好，↓越低越好）。

受试者	低层		高层				
	像素相关性↑	结构相似性指数↑	AlexNet(2)↑	AlexNet(5)↑	Inception↑	CLIP↑	SwAV↓
1	0.193	0.317	0.816	0.886	0.761	0.771	0.568
2	0.190	0.346	0.845	0.919	0.789	0.821	0.551
3	0.191	0.330	0.834	0.903	0.758	0.827	0.559
4	0.183	0.334	0.825	0.905	0.805	0.840	0.535
5	0.176	0.326	0.825	0.903	0.720	0.795	0.561
6	0.191	0.326	0.833	0.907	0.791	0.817	0.552
7	0.166	0.337	0.831	0.910	0.765	0.794	0.556
8	0.207	0.365	0.861	0.918	0.815	0.827	0.528
9	0.183	0.348	0.838	0.904	0.792	0.797	0.535
10	0.190	0.365	0.848	0.921	0.830	0.854	0.521
Ave	0.187	0.339	0.836	0.908	0.783	0.814	0.547

表12：使用H14+变分自编码器设置时各受试者的重建指标（ ↑ 越高越好，↓越低越好）。

	低层		高层				
受试者	像素相关性↑	结构相似性指数↑	AlexNet(2)↑	AlexNet(5)↑	Inception↑	CLIP↑	SwAV↓
1	0.156	0.301	0.755	0.762	0.653	0.646	0.658
2	0.175	0.323	0.801	0.852	0.643	0.743	0.608
3	0.171	0.290	0.793	0.853	0.651	0.730	0.621
4	0.167	0.307	0.798	0.851	0.707	0.761	0.589
5	0.174	0.295	0.783	0.847	0.670	0.723	0.611
6	0.174	0.304	0.806	0.845	0.669	0.720	0.612
7	0.154	0.315	0.764	0.828	0.655	0.711	0.622
8	0.196	0.335	0.817	0.859	0.691	0.734	0.590
9	0.166	0.315	0.765	0.827	0.678	0.701	0.605
10	0.194	0.337	0.810	0.856	0.703	0.746	0.594
Ave	0.173	0.312	0.789	0.838	0.672	0.721	0.611

表13：使用 H14+B32+变分自编码器设置时各受试者的重建指标（ ↑ 越高越好，↓越低越好）。

	低层		高层				
受试者	像素相关性↑	结构相似性指数↑	AlexNet(2)↑	AlexNet(5)↑	Inception↑	CLIP↑	SwAV↓
1	0.193	0.332	0.835	0.883	0.727	0.757	0.578
2	0.188	0.341	0.846	0.901	0.769	0.807	0.559
3	0.196	0.324	0.834	0.900	0.755	0.826	0.566
4	0.187	0.320	0.821	0.903	0.774	0.824	0.553
5	0.179	0.317	0.831	0.893	0.705	0.797	0.565
6	0.211	0.329	0.852	0.920	0.758	0.811	0.561
7	0.179	0.336	0.833	0.914	0.754	0.805	0.554
8	0.219	0.356	0.876	0.926	0.788	0.827	0.531
9	0.198	0.342	0.842	0.889	0.757	0.783	0.546
10	0.203	0.358	0.858	0.922	0.777	0.846	0.530
Ave	0.195	0.336	0.843	0.905	0.756	0.808	0.554

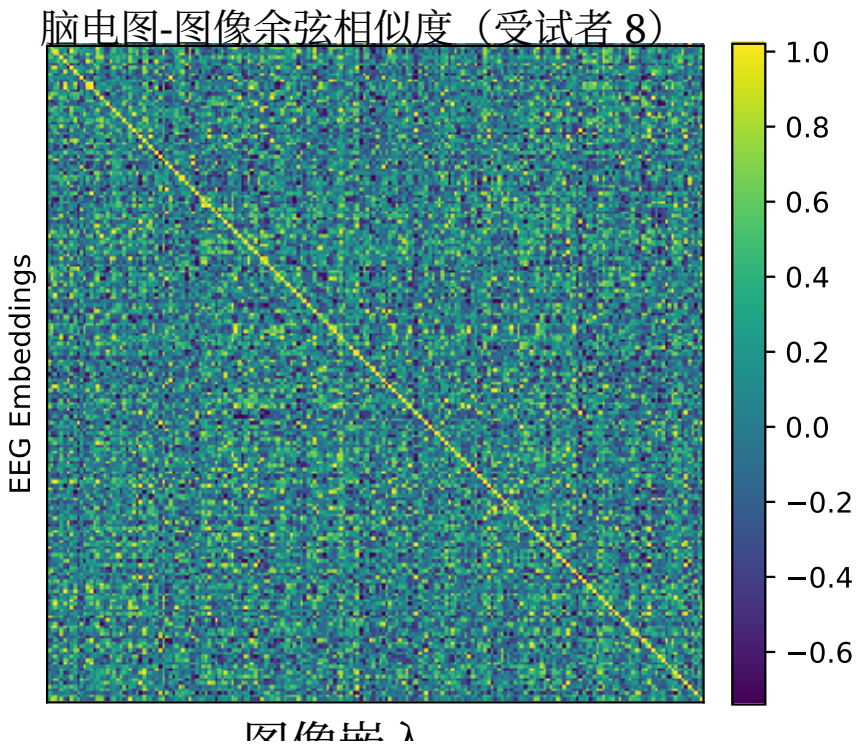


图 7：受试者 8 在 THINGS-EEG 数据集测试集划分上的脑电图嵌入与图像嵌入之间的余弦相似度矩阵。行对应脑电图嵌入，列对应图像嵌入；明显的对角线模式表明，每个脑电图嵌入都与其配对的图像嵌入最为相似。

D 补充消融研究

扩展视觉编码器消融实验 图 8 和表 14 对比了其他视觉编码器（DINO（Oquab 等，2023）和 SynCLR（Tian 等，2024））及其多编码器变体。在不同骨干网络上，加入 SDXL 变分自编码器都能持续提升受试者内和受试者间设置下的前1和前5检索准确率，其中 DINO 的提升尤为显著。将 DINO 与 CLIP 编码器（RN50 或 B32）配对可带来适度提升，而基于变分自编码器的组合增益最大，尤其是与 B32 融合时。对于 SynCLR，仅与 RN50 或 B32 堆叠时性能基本趋于饱和，并且可能略微损害受试者间迁移；而 SynCLR+变分自编码器和基于变分自编码器的三重堆叠取得了最佳总体结果。这些趋势表明，低层变分自编码器特征能够系统性地补充高层表示，从而提升不同视觉骨干网络上的脑电图-图像对齐效果。

脑电图特征选择对通道的影响 除非另有说明，我们使用刺激起始后 0–1000 毫秒的脑电图数据，并按照先前工作（Song 等，2024）将输入限制为 17 个枕叶–顶叶（O+P[2]）通道。表 15 表明，对于每种视觉编码器配置，与仅使用枕叶通道、仅使用顶叶通道、“其他”非视觉通道或完整的 63 通道电极布局相比，O+P 子集始终取得最高的 200 类检索准确率。

脑电图特征选择对时间窗的影响图9绘制了不同视觉编码器配置下，前1检索准确率随 EEG 时间窗的变化情况。对于累积时间窗 $[0, t]$，在完整的 RN50+B32+VAE 融合设置下，准确率从 0–100 ms 时的 21.8% 上升至 0–300 ms 时的 73.9%，随后保持在接近 0–1000 ms 基线（ 75.7% ）的水平；而排除早期活动的尾部时间窗 $[t, 1000]$ 则使准确率从 75.7%（0–1000 ms）降至 69.9%（100–1000 ms）

[2]P7,P5,P3,P1,Pz,P2,P4,P6,P8,PO7,PO3,POz,PO4,PO8,O1,Oz,O2

表14：扩展视觉编码器设置下，THINGS-EEG 上 200类零样本检索的前1和前5准确率（%）。前两列报告受试者内性能，后两列报告受试者间性能。结果为相同训练配置下全部 10名受试者的平均值，仅视觉编码器堆栈有所不同。括号中的数字表示相较于对应单编码器基线的绝对增益。

视觉设置	受试者内		受试者间	
	前1	前5	前1	前5
DINO	23.5	50.9	10.7	22.8
SynCLR	68.0	91.8	19.5	42.0
DINO+RN50	28.6 (+5.1)	57.7 (+6.8)	11.9 (+1.2)	26.3 (+3.5)
DINO+B32	35.7 (+12.2)	66.8 (+15.9)	13.3 (+2.6)	29.9 (+7.1)
DINO+变分自编码器	50.1 (**+26.6**)	79.9 (**+29.0**)	16.1 (**+5.4**)	34.8 (**+12.0**)
SynCLR+RN50	69.1 (+1.1)	91.9 (+0.1)	19.2 (-0.3)	41.0 (-1.0)
SynCLR+B32	68.4 (+0.4)	92.4 (+0.6)	19.2 (-0.3)	42.3 (+0.3)
SynCLR+变分自编码器	78.5 (**+10.5**)	96.4 (**+4.6**)	23.0 (**+3.5**)	46.5 (**+4.5**)
DINO+RN50+变分自编码器	53.3 (+29.8)	82.6 (+31.7)	17.3 (+6.6)	35.8 (+13.0)
DINO+B32+变分自编码器	58.5 (**+35.0**)	85.9 (**+35.0**)	19.3 (**+8.6**)	39.8 (**+17.0**)
SynCLR+RN50+变分自编码器	78.7 (+10.7)	96.5 (+4.7)	23.4 (+3.9)	47.1 (+5.1)
SynCLR+B32+变分自编码器	78.9 (**+10.9**)	96.5 (**+4.7**)	23.4 (**+3.9**)	47.2 (**+5.2**)

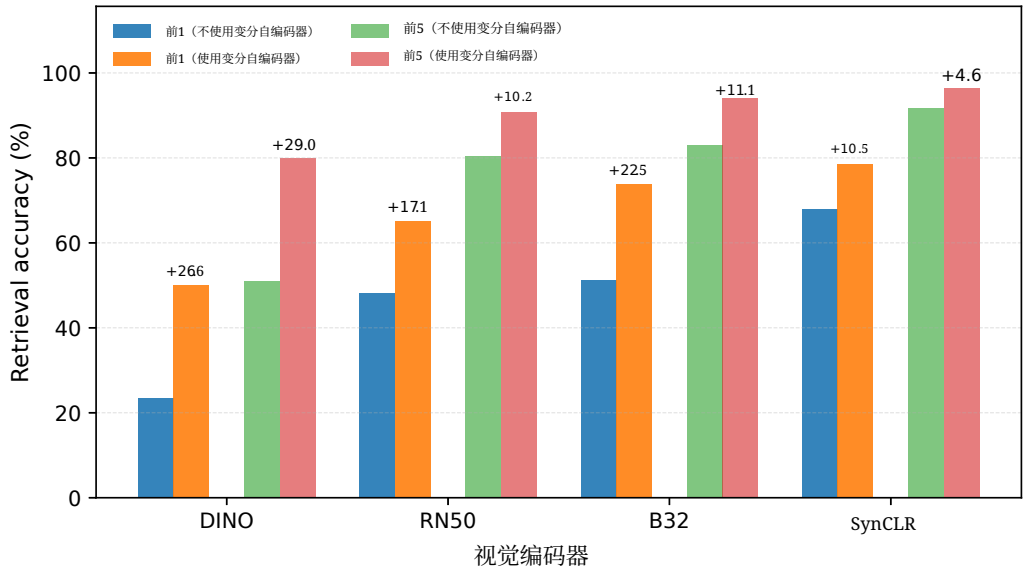


图8：将脑电图与单个编码器对齐或与编码器+变分自编码器混合对齐时，在 THINGS-EEG 上进行 200 类零样本检索的前1和前5准确率（%）。柱状图展示了各设置下的表现，以及加入 SDXL VAE 后的绝对提升；结果为全部10名受试者的平均值。

以及 13.7%（300–1000 毫秒），在 500–1000 毫秒时（1.7%）接近随机水平。比较不同编码器可见，在最早的累积时间窗中，基于VAE的配置往往优于基于CLIP的配置；而在较晚的尾部时间窗中，基于CLIP的配置优于VAE，这表明早期脑电活动可能更贴合低层的类像素特征，而后期活动可能更贴合高层语义特征。对于更完整的时间窗（即 [0, t]较大且 [t 1000]较小），以CLIP为主的配置和融合配置通常优于单独使用VAE。这与一种可能性相符：几百毫秒后，对200类零样本检索最有用的脑电信号主要由高层视觉信息主导。在所有

表 15：不同视觉编码器设置（行）和通道选择（列）下，THINGS-EEG 上 200类零样本检索的前1和前5准确率（%）。所有数值均为相同训练配置下 10名受试者的平均值，仅所选通道有所不同。

	枕叶		顶叶		O+P（我们的方法）		其他		All	
方法	前1	前5	前1	前5	前1	前5	前1	前5		
RN50	43.1	75.51	21.8	52.1	48.1	80.5	9.4	26.9	43.1	75.7
B32	48.3	80.0	23.9	52.8	51.2	83.0	8.5	26.4	44.1	78.3
VAE	45.9	76.6	10.0	27.7	44.1	75.5	2.8	10.9	27.0	59.0
RN50+B32	51.9	83.2	25.6	54.7	56.2	86.2	8.7	27.2	47.4	80.8
RN50+变分自编码器	65.1	90.3	19.8	47.1	65.2	90.7	5.9	18.4	43.6	76.8
B32+变分自编码器	71.3	93.1	26.2	57.1	73.7	94.1	9.3	27.2	59.7	88.9
RN50+B32+变分自编码器	72.6	93.5	27.4	58.9	75.7	94.6	9.1	28.4	62.5	90.0

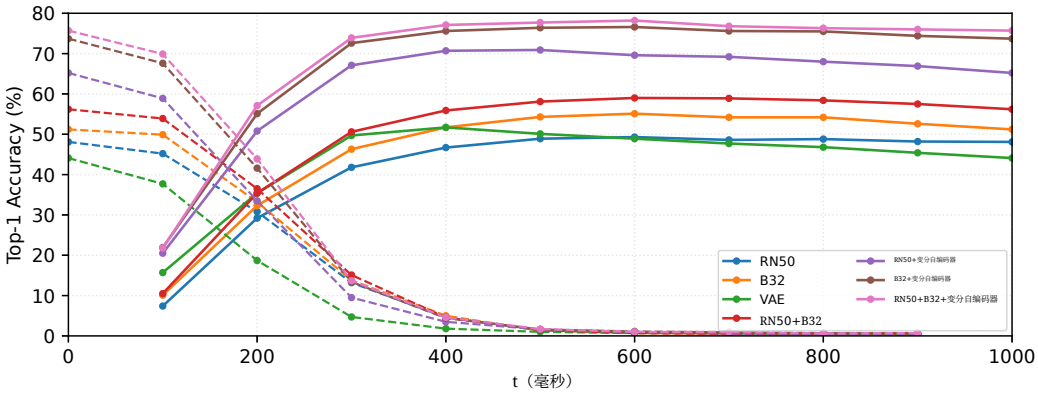


图 9：不同视觉编码器配置下，Top-1 检索准确率（%）随 EEG 时间窗变化的情况。实线表示累积时间窗 $[0, t]$；虚线表示尾部时间窗$[t, 1000]$，其中 t=0 表示完整的 $[0, 1000]$ 毫秒区间。

时间窗中，完整的 RN50+B32+VAE 融合始终取得最佳表现，进一步证实了我们方法的有效性。

编码器掩蔽分析 为探究 RN50+B32+VAE 融合模型中各视觉分支的相对贡献，我们保持训练好的模型不变，仅在推理时对单个编码器嵌入进行零掩蔽（表16）。移除 VAE 或 B32 分支会导致受试者内和受试者间准确率均显著下降，而掩蔽 RN50 的影响相对较小，这表明在此融合配置中，VAE 和 B32 提供了主要的互补信号。

表 16：RN50+B32+VAE 融合设置下，脑到图像检索的编码器掩蔽消融实验。在相同训练配置下，推理时屏蔽一个视觉编码器，并报告 THINGS-EEG 上 200 类零样本检索的平均前1/前5准确率（%）。

设置	受试者内		受试者间	
	前1	前5	前1	前5
不使用 RN50	70.6	92.7	19.8	41.2
不使用 B32	41.2	73.2	11.9	27.4
不使用变分自编码器	29.6	60.2	9.3	23.4
FULL	**75.7**	**94.6**	**20.0**	**44.1**