

# DreamDiffusion：具有时间掩蔽信号建模和 CLIP 对齐的高质量 EEG 到图像生成

Yunpeng Bai<sup>1,5</sup> , Xintao Wang<sup>3</sup>, Yan-Pei Cao<sup>4</sup>, Yixiao Ge<sup>2</sup>, Chun Yuan<sup>1</sup> , and Ying Shan<sup>2</sup>

1 清华大学深圳国际研究生院，中国  
2 ARC Lab, 腾讯PCG, 3 快手科技, 4 VAST5 德克萨斯大学奥斯汀分校  
<https://github.com/bbaaii/DreamDiffusion>

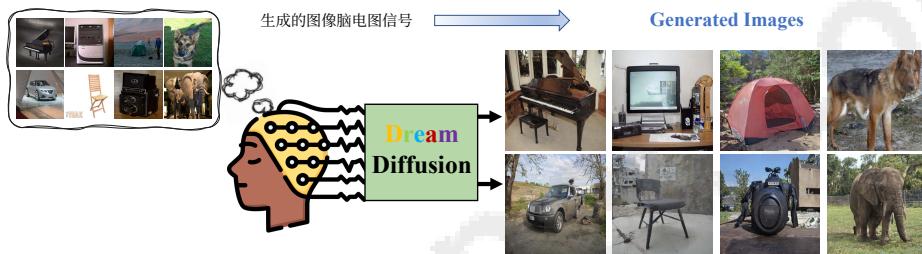


图 1：我们提出的 DreamDiffusion 能够直接从大脑脑电图 (EEG) 信号生成高质量图像，无需将思想转化为文本。

抽象的。本文介绍了 DreamDiffusion，这是一种直接从大脑脑电图 (EEG) 信号生成高质量图像的新方法，无需将思想转化为文本。DreamDiffusion 利用预先训练的文本到图像模型，并采用时间掩蔽信号建模来预训练 EEG 编码器，以实现有效且稳健的 EEG 表示。此外，该方法进一步利用 CLIP 图像编码器提供额外的监督，以更好地将脑电图、文本和图像嵌入与有限的脑电图图像对齐。总的来说，所提出的方法克服了使用脑电信号进行图像生成的挑战，例如噪声、信息有限和个体差异，并取得了有希望的结果。定量和定性结果证明了该方法的有效性，是朝着便携式和低成本“想到图像”迈出的重要一步，在神经科学和计算机视觉方面具有潜在的应用。

关键词：图像生成·脑电解码·模型预训练

 通讯作者。

## 1 简介

近年来，图像生成[4,15,21]取得了长足的进步，特别是在文本到图像生成方面取得了突破[1,12,29,30,33]。最近的文本到图像的生成使得人们的想法能够被创造成由文本控制的精美绘画和艺术品。我们很好奇是否可以直接通过大脑活动（例如脑电图（EEG）记录）来控制图像创建，而无需在创建之前将我们的想法转化为文本。这种“想到图像”的前景广阔，可以例如，它可以极大地提高艺术创作的效率，并帮助捕捉那些转瞬即逝的灵感。甚至可能有助于心理治疗，有可能帮助患有自闭症和语言障碍的儿童。最近的一些作品，例如 MinD-Vis [7] 和 [39]，尝试基于 fMRI（功能性磁共振成像）信号重建视觉信息，这是另一种测量大脑活动的方法，他们证明了从大脑活动中重建高质量结果的可行性，但是，它们离我们利用大脑信号方便高效地创造的目标还很遥远。1) 由于fMRI设备不便于携带，需要专业人员操作，因此捕获fMRI信号比较困难。2) MRI 采集成本高。它们极大地阻碍了这种方法在实际艺术生成中的广泛使用。相比之下，EEG（脑电图）是一种非侵入性且低成本的记录大脑电活动的方法。便携式商业产品现在可以方便地采集EEG信号，为未来的艺术生成展现了巨大的潜力。在这项工作中，我们旨在利用预先训练的文本到图像模型（即稳定扩散[31]）的强大生成能力，直接从大脑脑电图信号生成高质量图像。然而，这并非易事，并且存在两个挑战。1) 脑电图信号是非侵入性捕获的，因此本质上是有噪声的。此外，脑电图数据有限，个体差异不容忽视。如何从具有如此多约束的脑电信号中获得有效且鲁棒的语义表示？2) 由于使用了CLIP[27]以及对大量文本-图像对的训练，稳定扩散中的文本和图像空间很好地对齐。然而，脑电信号有其自身的特点，其空间与文本、图像有很大不同。如何将脑电图、文本和图像空间与有限且嘈杂的脑电图图像对齐？为了解决第一个挑战，我们建议使用大量脑电图数据而不是仅使用稀有的脑电图图像对来训练脑电图表示。具体来说，我们采用屏蔽信号建模来根据上下文线索预测丢失的标记。与MAE[17]和MinD-Vis[7]将输入视为二维图像并掩盖空间信息不同，我们考虑EEG信号的时间特征，并深入挖掘人们大脑时间变化背后的语义。我们随机屏蔽一部分令牌，然后在时域中重建这些屏蔽的令牌。这样，预训练的 reconstruct those masked ones *in the time domain*. In this way, the pre-trained

编码器学习对不同人和各种大脑活动的脑电图数据的深入理解。至于第二个挑战，以前的方法[7,39]通常使用少量噪声数据对直接微调稳定扩散（SD）模型。然而，仅使用最终的图像重建损失来通过端到端微调 SD 很难学习大脑信号（例如，EEG 和 fMRI）与文本空间之间的准确对齐。因此，我们建议采用额外的 CLIP [27] 监督来协助脑电图、文本和图像空间的对齐。具体来说，SD 本身使用CLIP的文本编码器来生成文本嵌入，这与前一阶段的屏蔽预训练EEG嵌入有很大不同。我们利用 CLIP 的图像编码器来提取与 CLIP 文本嵌入很好地对齐的丰富图像嵌入。然后使用这些 CLIP 图像嵌入来进一步优化 EEG 嵌入表示。因此，精细化的 EEG特征嵌入可以很好地与CLIP图像和文本嵌入对齐，并且更适合SD图像生成，从而提高生成图像的质量。配备上述两种精致的设计，我们提出的方法，即DreamDiffusion，可以从脑电图信号生成高质量和逼真的图像。我们的贡献可总结如下。1) 我们提出了 DreamD-iffusion，它利用强大的预训练文本到图像扩散模型仅从脑电图信号生成逼真的图像。这是朝着便携式和低成本的“想到图像”迈出的又一步。2) 我们专门探索了针对脑电图数据的时间掩蔽信号建模方法，以学习有效的表示，这对于后续的相关工作很有用。3) 我们进一步利用 CLIP 图像编码器提供额外的监督，以更好地将 EEG、文本和图像嵌入与有限的 EEG-图像对对齐。4) 定量和定性结果表明了我们方法的有效性。

## 2 相关作品

### 2.1 从大脑活动生成图像

使用大脑信号（包括功能磁共振成像和脑电图）来生成图像一直是一个活跃的研究领域。对于功能磁共振成像的使用，传统方法依靠功能磁共振成像-图像配对数据来训练模型以预测功能磁共振成像的图像特征。这些图像特征将被输入 GAN [35]，以在测试期间进行刺激重建。然而，最近的研究 [3] 提出了无监督方法，例如可重构自动编码器设计，从不成对的 fMRI 和图像中学习，并利用回归模型 [24,26] 提取潜在的 fMRI 表示，该表示可用于微调预编码器。-训练条件 BigGAN [5] 用于解码。最近的工作 MinD-Vis [8] 集成了 SC-MBM 和 DC-LDM，以生成更可信的图像，并保留更好的语义信息。类似地，还探索了使用深度学习技术从 EEG 信号生成图像。Brain2image [22] 使用 LSTM 和生成方法来学习更紧凑的脑电图数据表示，以生成引起特定大脑反应的视觉刺激。ThoughtViz [40] 采取

erating visual stimuli that evoke specific brain responses. ThoughtViz [40] takes

编码脑电图信号作为输入来生成相应的图像，即使训练数据有限。[9]使用脑电图作为监督信号来学习语义特征表示并实现与语义图像编辑相当的性能。

## 2.2 Model pre-training

预训练模型在计算机视觉领域变得越来越流行，各种自监督学习方法专注于不同的预训练任务[13,25,42]。这些方法通常利用借口任务，例如对比学习 [2, 16]（对图像相似性和相异性进行建模）或自动编码 [6]（从屏蔽部分恢复原始数据）。特别是，掩蔽信号建模 (MSM) 通过从视觉信号的高掩蔽比 [17, 43] 和自然语言的低掩蔽比 [10, 28] 恢复原始数据，成功地学习了下游任务的有用上下文知识。。另一种最近的方法是 CLIP [27]，它通过对从互联网上的各种来源收集的 4 亿个文本图像对进行预训练来构建多模态嵌入空间。CLIP 学习到的表示非常强大，可以在多个数据集上进行最先进的零样本图像分类，并提供一种估计文本和图像之间语义相似性的方法。

## 2.3 扩散模型

扩散模型作为生产高质量内容的生成模型变得越来越流行[36]。扩散模型的基本形式是由双向马尔可夫状态链定义的概率模型[18]。这些模型[11,18,32,38]由于它们与图像数据的归纳偏差的自然契合而表现出强大的生成能力。最佳合成质量通常是在训练期间使用重新加权的目标时实现的[18]，允许在图像质量和压缩能力之间进行权衡。然而，在像素空间中评估和优化这些模型在计算上既昂贵又耗时[19,23,34,37,41]。为了应对这些挑战，一些扩散模型适用于较低维度的压缩潜在空间，例如所提出的 LDM [31]。通过使用 KL 正则化自动编码器将图像压缩为低维潜在特征，然后使用相同的潜在空间特征重建它们，LDM 降低了计算成本，同时保持了合成质量。

## 3 提出的方法

我们的方法包括三个主要组成部分：1) 针对有效且鲁棒的脑电图编码器进行掩蔽信号预训练，2) 通过预先训练的稳定扩散对有限的脑电图图像对进行微调，以及3) 对齐脑电图、文本和图像。使用 CLIP 编码器的年龄空间。首先，我们利用掩蔽信号建模  
age spaces using CLIP encoders. Firstly, we leverage masked signal modeling

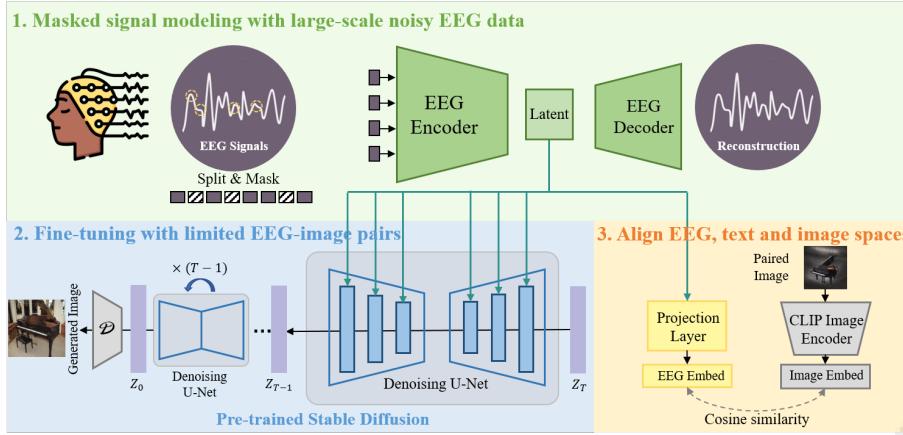


图 2: DreamDiffusion 概述。我们的方法包括三个主要组成部分: 1) 对有效且鲁棒的脑电图编码器进行掩蔽信号预训练, 2) 通过预训练的稳定扩散对有限的脑电图图像对进行微调, 3) 对齐脑电图、文本和使用 CLIP 编码器的图像空间。

使用大量嘈杂的脑电图数据来训练脑电图编码器来提取上下文知识。然后使用所得的脑电图编码器通过交叉注意机制为稳定扩散提供条件特征。为了增强脑电图特征与稳定扩散的兼容性, 我们通过在微调过程中减小脑电图嵌入和 CLIP 图像嵌入之间的距离来进一步对齐脑电图、文本和图像嵌入空间。之后, 我们获得了 DreamDiffusion, 它能够仅从 EEG 信号生成高质量图像。

### 3.1 掩蔽信号预训练以实现有效且鲁棒的脑电图表示

EEG (脑电图) 数据是人脑产生的电活动的记录, 使用放置在头皮上的电极进行测量。它是一种非侵入性且低成本的测量大脑活动的方法。脑电图数据有几个特点。首先, 数据是二维的, 一维代表放置在头皮上的通道或电极, 另一维代表时间。脑电图的时间分辨率较高, 这意味着它可以捕获毫秒量级的大脑活动快速变化。但是, 脑电图的空间分辨率较低, 这意味着它很难精确定位活动的来源。其次, 脑电图信号变化很大, 受到年龄、睡眠和认知状态等因素的影响。最后, 脑电图数据通常充满噪音, 需要仔细处理和分析才能提取有意义的信息。由于脑电图数据固有的变异性和噪音, 传统建模方法通常很难从脑电图信号中提取有意义的信息。

ing methods often struggle to extract meaningful information from EEG signals.

脑电图数据本身包含许多与相应视觉概念（噪声）无关的信号，使得通过未经预训练的编码器简单地利用脑电图的结果不准确。因此，采用掩蔽信号建模技术已被证明可以有效地从噪声和可变数据中捕获上下文信息[7, 17]，这代表了从大规模噪声脑电图数据中获取有意义的上下文知识的有前途的途径。与MAE[17]和MinD-Vis[7]将输入视为二维图像并掩盖空间信息不同，我们考虑EEG信号的时间特征，并深入挖掘人们的时间变化背后的语义。鉴于脑电信号的时间分辨率较高，我们首先将其在时域上划分为tokens，并随机屏蔽一定比例的tokens。随后，这些tokens将被转化为使用一维卷积层进行嵌入。然后，我们使用非对称架构（例如MAE[17]）根据周围标记的上下文线索来预测丢失的标记。通过重建掩蔽信号，预先训练的脑电图编码器可以深入了解不同人和各种大脑活动的脑电图数据。

### 3.2 在有限的脑电图图像上使用稳定扩散进行微调

从掩蔽信号预训练中获得脑电图信号的有效表示后，我们利用预训练的稳定扩散（SD）模型来生成图像。稳定扩散涉及逐渐对正态分布变量进行去噪以学习数据分布。SD增强了交叉注意机制，以实现更灵活的条件图像生成，最常见的条件是文本提示。稳定扩散在从各种类型的信号（例如标签、文本和语义图）生成高质量图像方面表现出了强大的生成能力。稳定扩散在潜在空间上运行。给定像素空间中的图像 $x$ ， $x$ 由VQ编码器 $E(\cdot)$ 编码以获得相应的latent $z = E(x)$ 。条件信号是由UNet中的交叉注意力机制引入的。这种交叉注意力还可以合并来自脑电图数据的条件信息。具体来说，EEG编码器 $y$ 的输出进一步用投影仪 $\tau_\theta$ 投影到嵌入 $\tau_\theta(y) \in \mathbb{R}^{d\tau}$ 中。然后，这个EEG表示通过交叉注意力层合并到U-Net中，实现 $\text{Attention}(Q, K, V) = \text{softmax}(QK^T)V$

$$Q = W_Q^{(i)} \cdot \varphi_i(z_t), K = W_K^{(i)} \cdot \tau_\theta(y), V = W_V^{(i)} \cdot \tau_\theta(y), \quad (1)$$

其中 $\varphi_i(z_t) \in \mathbb{R}^{d\tau \times d\tau}$ 表示UNet的中间值。瓦 $W_V^{(i)} \in \mathbb{R}^{d\tau \times d\tau}$ 是具有可学习参数的投影矩阵。在微调过程中，我们一起优化了UNet的EEG编码器和交叉注意头。我们保持StableDiffusion的其余部分固定。我们使用以下SD损失函数进行微调。

Diffusion fixed. We use the following SD loss function for fine-tuning.

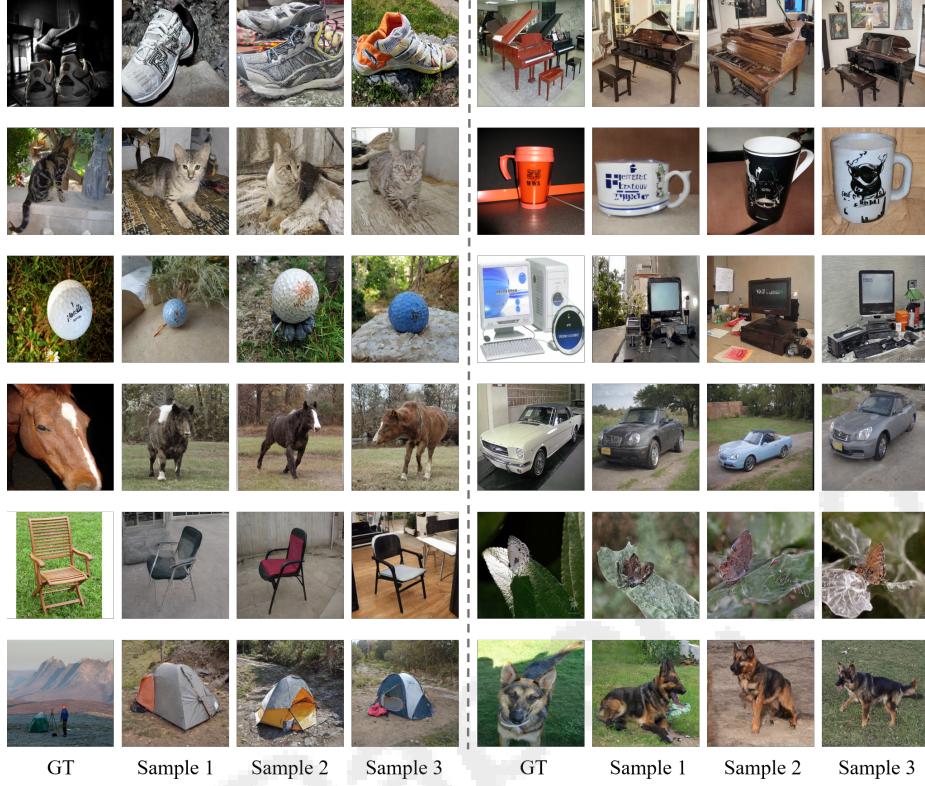


图 3：主要结果。左边的图像描绘了配对图像数据，而右边的三幅图像代表采样结果。可以看出，我们的模型从脑电图数据中生成了高质量的图像，并且这些图像与脑电图数据准确匹配。

$$L_{SD} = \mathbb{E}_{x, \epsilon \sim \mathcal{N}(0,1), t} \left[ \|\epsilon - \epsilon_\theta(x_t, t, \tau_\theta(y))\|_2^2 \right], \quad (2)$$

其中  $\epsilon_\theta$  是作为 UNet 实现的去噪函数。

### 3.3 将 EEG、文本和图像空间与 CLIP 编码器对齐

接下来，我们将对预训练获得的脑电表示进行微调，使其更适合生成图像。预训练的稳定扩散模型专门针对文本到图像的生成进行训练；然而，脑电信号有其自身的特点，其潜在空间与文本和图像有很大不同。因此，使用有限的脑电图图像配对数据直接微调稳定扩散模型不太可能准确地将脑电图特征与文本嵌入对齐。

features with the text embeddings.



图 4：与 Brain2Image 的比较。DreamDiffusion 生成的图像质量明显高于 Brain2Image 生成的图像。

由于使用 CLIP [27] 以及对大量文本图像对的训练，稳定扩散中的文本和图像空间很好地对齐。因此，我们建议采用额外的 CLIP [27] 监督来协助对齐脑电图、文本和图像空间。具体来说，从预训练编码器获得的 EEG 特征通过投影层转换为与 CLIP 相同维度的嵌入。然后，我们使用损失函数来最小化 EEG 嵌入与从 CLIP 图像编码器获得的图像嵌入之间的距离。CLIP 模型在微调过程中是固定的。损失函数定义如下：

$$\mathcal{L}_{clip} = 1 - \frac{E_I(I) \cdot h(\tau_\theta(y))}{|E_I(I)| |h(\tau_\theta(y))|}, \quad (3)$$

其中  $h$  是投影层， $E_I$  是 CLIP 图像编码器。这种损失函数可以促使脑电图特征与图像更加紧密地对齐，从而与文本特征更加相似。这样，我们就可以将脑电信号、文本和图像排列在一个统一的空间中。优化后的脑电图嵌入表示更适合 SD 图像生成，从而提高了生成图像的质量。CLIP 对齐主要不是为了增强对 EEG 的理解，而是为了提高从预训练获得的有效 EEG 表示对稳定扩散的适应性。

EEG representations obtained from pretraining to Stable Diffusion.

## 4 实验与分析

### 4.1 实施细节

脑电图表示预训练的数据。我们在 MOABB [20] 平台上收集了来自 400 多名受试者的约 120,000 个通道范围从 30 到 128 的脑电图数据样本，用于脑电图预训练。这些数据包括平台中 30 多个通道的所有脑电数据。MOABB 是一个软件包，旨在通过提供通用格式的公开脑电图数据集以及一套最先进的算法来促进脑机接口 (BCI) 算法的开发。该平台使研究人员能够轻松地使用自动统计分析验证新算法，无需进行耗时且不可靠的数据预处理。这些数据包含各种各样的脑电图数据，包括查看物体、运动想象和观看视频等任务。我们的目标是从不同的脑电图数据中学习通用表示，而不对脑电图数据的类型有特定要求。由于用于数据采集的设备的变化，这些脑电图数据样本的通道数存在显着差异。为了便于预训练，我们通过用复制值填充缺失的通道，将所有通道数较少的数据统一填充到 128 个通道。在预训练过程中，每 4 个相邻时间步被分组为一个 token，每个 token 通过投影层转换为 1024 维嵌入，用于后续的掩蔽信号建模。损失函数计算重建的 EEG 信号和原始 EEG 信号之间的 MSE。重建是在整个 128 个通道集上执行的，而不是在每个通道的基础上执行。损失仅在屏蔽补丁上计算。重建是在整个 128 个通道集上执行的，而不是在每个通道的基础上执行。解码器在预训练后被丢弃。我们的数据处理方法包括 Beta (15-31 Hz) 和 Gamma (32-70 Hz) 频段，因为它们传达涉及视觉感知的信息以及潜在有用的频率。低于 5Hz 的 EEG 信号通常只出现在深度睡眠期间，因此我们使用的数据中不太可能出现低于 5Hz 的有用信号。由于预训练中使用的数据来自不同地区和设备，具有不同的线路噪声，因此我们没有考虑这一点进行统一处理。此外，根据之前工作的实验，线路噪声的影响并不显着。配对脑电图图像数据。我们采用 ImageNet-EEG [22] 数据集进行“想法到图像”实验，该数据集是从 6 名受试者获得的脑电图记录的集合，同时向他们展示了 ImageNet 数据集中属于 40 个不同对象类别的 2000 张图像。每个类别由 50 张图像组成，每张图像呈现 0.5 秒，每 50 张图像暂停 10 秒。使用 128 通道 Brainvision EEG 系统记录 EEG 数据，总共产生 12000 个 128 通道 EEG 序列。该数据集包括各种物体的图像，例如动物（狗、猫、大象等）、车辆（飞机、自行车、汽车等）以及日常生活

(dogs, cats, elephants, etc.), vehicles (airliners, bikes, cars, etc.), and everyday

物体（电脑、椅子、杯子等）。更多细节可以参见相关参考文献 [22]。其他细节。我们使用稳定扩散 1.5 版本来生成图像。EEG 信号的掩模比设置为 75%。所有 EEG 信号均在 5-95 Hz 频率范围内进行过滤。随后，信号被截断为通用长度 512。编码器经过 500 个 epoch 的预训练，并通过稳定扩散进行另外 300 个周期的微调。EEG 的预训练模型类似于 ViT-Large [14]。训练和测试是在同一主题上进行的，论文中呈现的所有结果都是使用主题 4 的数据生成的。有关更多结果，请参阅补充材料。

表1：更多比较评估。在比较各种定量指标的背景下，我们的方法全面显着优于以前的方法。

Methods	FID ↓	IS ↑	PSNR ↑	SSIM ↑	LPIPS ↓
Brain2Image	18.76	5.06	12.8	0.213	0.701
Ours	3.61	28.54	14.6	0.267	0.644

#### 4.2 与 Brain2Image 的比较

在本节中，我们将我们提出的方法与 Brain2Image [22] 进行比较。但是，它仅提供少数类别的结果，并且不提供参考实现。鉴于此，我们对 Brain2Image 论文中展示的几个类别（即客机、杰克灯笼和熊猫）的结果进行了定性比较。为了确保公平比较，我们遵循与 Brain2Image 概述相同的主观评估策略，并在图 4 中展示了不同方法的生成实例。顶行描述了 Brain2Image 生成的结果，而底行是由我们提出的方法 DreamDiffusion 生成的。我们观察到 DreamDiffusion 生成的图像质量明显高于 Brain2Image 生成的图像，从而验证了我们提出的方法的有效性。由于可用图像数量有限，使用 FID/IS 等指标可能不稳定，无法有效衡量质量因此，我们在这里提供 FID/IS 指标仅供参考（表 1）。指标是使用 Brain2Image 论文中的图像计算的。尽管如此，我们的方法在这些定量指标方面显着优于以前的方法。我们还在这里添加了一些其他基于相似性的指标或基于感知的指标。由于我们的方法的目的不是精确的图像重建，而是从脑电图信号生成相应的概念图像，因此这些指标仅供参考。

provided for reference only.

## 4.3 消融研究

在本节中，我们使用各种案例对所提出的框架进行了多项消融研究。我们通过采用 50 路 top-1 精度分类任务来评估不同方法的有效性。我们使用预先训练的 ImageNet1K 分类器 [14] 来确定生成图像的语义正确性。真实图像和生成的图像都将被输入到分类器中。然后，我们将验证生成图像的 top-1 分类是否与 50 个选定类别中的真实分类相匹配。只要生成图像的语义分类结果与地面实况一致，就认为生成图像是正确的。

表 2：消融研究的定量结果。E 和 A 分别代表编码器和交叉注意力头的微调。Params: EEG 编码器中的可训练参数。第 1-4 行主要是为了说明没有预训练的剪辑微调对结果的影响。第 5-7 行旨在解释不同掩模比率对结果的影响。第 8-11 行演示了编码器参数量的影响，其中较大的参数不一定会产生更好的结果。最后，第 12-14 行阐明了分别微调编码器或注意力层的效果。

Model	MSM Pretraining	CLIP Finetuning	Mask Ratio	E + A	Params	Acc (%)
Full	✓	✓	0.75	E + A	<b>297M</b>	<b>45.8</b>
1	✗	✗	-	E + A	<b>297M</b>	4.2
2	✗	✗	-	E + A	<b>18.3M</b>	3.7
3	✗	✓	-	E + A	<b>297M</b>	32.3
4	✗	✓	-	E + A	<b>18.3M</b>	24.5
5	✓	✓	0.25	E + A	<b>297M</b>	19.7
6	✓	✓	0.5	E + A	<b>297M</b>	38.3
7	✓	✓	0.85	E + A	<b>297M</b>	33.4
8	✓	✓	0.75	E + A	<b>458M</b>	38.5
9	✓	✓	0.75	E + A	<b>162M</b>	36.6
10	✓	✓	0.75	E + A	<b>74M</b>	29.8
11	✓	✓	0.75	E + A	<b>18.3M</b>	28.7
12	✓	✓	0.75	E only	<b>297M</b>	22.4
13	✓	✗	0.75	E + A	<b>297M</b>	28.3
14	✓	✗	0.75	A only	<b>297M</b>	20.9

预训练的作用：为了证明大规模脑电图数据预训练的有效性，我们通过使用未经训练的编码器训练多个模型来进行验证。其中一个模型与完整模型相同，而另一个模型具有浅脑电图编码层，只有两层，以避免过度拟合数据。在训练过程中，两个模型分别在有剪辑监督和无剪辑监督的情况下进行训练，结果如表 2 的模型 1-4 所示。可以观察到，未经预训练的模型准确率有所下降。

decreased.

脑电图数据本身包含许多与相应视觉概念（噪声）无关的信号，使得通过未经预训练的编码器简单地利用脑电图的结果不准确。预训练方法通过随机掩蔽操作，在预训练过程中将模型暴露于各种类型的噪声和干扰中，以重建原始信号，从而使模型能够学习视觉概念的有效表示。使用预训练编码器后准确性的提高表现为“有效”和“稳健”。掩模比：我们研究确定使用 EEG 数据进行 MSM 预训练的最佳掩模比。如表 2 的模型 5-7 所示，过高或过低的掩模比都会对模型的性能产生不利影响。在掩模比为 0.75 时实现了最高的总体精度。这一发现很重要，因为它表明，与通常使用低掩蔽比的自然语言处理不同，在 EEG 上执行 MSM 时，高掩蔽比也是一个更好的选择。CLIP 对齐：我们方法的关键之一是对齐 EEG 表示通过 CLIP 编码器处理图像。CLIP 对齐主要不是为了增强对 EEG 的理解，而是为了提高预训练获得的有效 EEG 表征对稳定扩散的适应性。为了验证这种方法的有效性，我们进行了实验 13-14，如表 2 所示。可以观察到当不使用 CLIP 监督时，模型的性能显着下降。事实上，如图 5 右下角所示，即使在没有预训练的情况下，使用 CLIP 来对齐 EEG 特征仍然可以产生合理的结果，这凸显了 CLIP 监督在我们的方法中的重要性。其他方面：我们进一步说明了通过解释图 5 和表 2 来了解其余部分的作用。如图 5 右上角的两幅图像所示，使用未经预训练和 CLIP 微调的编码器会导致生成质量非常差。即使编码器经过训练但随后没有进行微调，我们发现它可以在一定程度上解码概念，但质量明显较差（底部第一）。如果不同时微调稳定扩散的交叉注意力层，就无法获得准确的结果（底部第二个）。在微调期间不使用 CLIP 解码相应的概念会产生一些准确但不完全精确的结果（底部第三个）。虽然单独使用 CLIP 进行微调可以在一定程度上使 EEG 与相应的概念保持一致，但其有效性不如完全预训练的方法（下 4）。总之，使用稳定扩散从 EEG 数据生成高质量图像并不简单，论文中讨论的方法的各个方面都是不可或缺的。查看表 2 中的结果，第 1-4 行主要旨在说明剪辑的影响无需对结果进行预训练即可进行微调。可以观察到，无论编码器中的参数数量如何，剪辑监督都有助于构建从脑电图信号到图像的映射，但它不一定有助于学习有效的脑电图表示。无论如何，与预训练的完整方法相比，准确性较低。第 8-11 行演示了编码器参数量的影响，表明较大的参数会影响编码器参数量。

the impact of encoder parameter volume, indicating that larger parameters do



图 5：消融研究的定性结果。右上角的两张图像说明了使用编码器而无需使用 CLIP 进行预训练和微调的效果。当编码器在训练后没有进一步微调时的结果（下行第一行）。当稳定扩散的交叉注意层不同时微调时的结果（下一行中的第二个）。微调过程中不使用 Clip 时的结果（下一行第 3 行），以及仅使用 Clip 进行微调时的结果（下一行第 4 行）。

并不总是能带来更好的结果。当参数量过大时，可能会拟合脑电信号中一些不相关的噪声。第 12-14 行旨在解释仅微调编码器或注意层的效果。可以看出，单独对其中任何一个进行微调都不利于脑电图的表示适应稳定扩散的条件输入，导致相应的精度下降。

#### 4.4 超越粗略的类别信息。

与重建相比，我们的方法旨在利用脑电图作为条件输入来生成图像。有时，我们的大脑会构思出抽象概念，而不是具体实体。与 Brain2Image 一样，我们的目标是使用包含抽象概念的脑电信号来生成相应的图像，而不是对实体进行精确重建。此外，当前实验中使用的数据主要包含类别级信息，因为在数据采集过程中每个图像显示 0.5 秒。未来我们将探索更详细的图像生成级别，例如收集具有丰富语义的脑电数据。虽然目前脑电数据在实验结果中仅提供类别级别的粗粒度信息，但我们的方法旨在探索可能的

level in experimental results currently, our method aims to explore the possi-



图 6: DreamDiffusion 的失败案例。某些特定类别由于形状或颜色相似而被错误地映射到其他类别。

从脑电图生成图像的能力，而不是仅仅用脑电图代替类别信息。如果我们使用类别标签代替CLIP作为监督，准确率将达到86.7%。如果直接使用类别标签作为输入，准确率可以达到97.2%。然而，添加类别标签并不是一个好的做法，因为我们的目标是将来使用语义上比类别信息更丰富的脑电图。利用CLIP结合脑电图像配对数据进行微调，无疑是未来应用更好、更合适的选择。

## 5 结论

本文提出了一种新方法，DreamDiffusion，用于从脑电图信号生成高质量图像，这是一种非侵入性且易于获得的大脑活动来源。该方法利用从大型脑电图数据集中学到的知识和图像扩散模型的强大生成能力，解决了与基于脑电图的图像生成相关的挑战。通过预训练和微调方案，脑电图数据可以被编码为适合的表示形式。使用稳定扩散生成图像。我们的方法代表了从大脑活动生成图像领域的重大进步。局限性。图 6 显示了一些失败案例，其中某些类别映射到具有相似形状或颜色的其他类别。我们认为这可能是由于人脑在识别物体时将形状和颜色视为两个重要因素。尽管如此，DreamDiffusion 具有广泛应用的潜力，例如神经科学、心理学和人机交互。

## Acknowledgements

This work was supported by the National Key R&D Program of China (2022YFB4701400/4701402), SSTIC Grant(KJZD20230923115106012, KJZD20230923114916032), and Beijing Key Lab of Networked Multimedia.

## References

1. Bai, Y., Wang, C., Xie, S., Dong, C., Yuan, C., Wang, Z.: Textir: 基于文本的可编辑图像恢复的简单框架。arXiv 预印本 arXiv: 2302.14736(2023)2。Becker, S., Hinton, G.E.: 发现随机点立体图中表面的自组织神经网络。《自然》355(6356), 161–163 (1992)3。Bird, C.M.、Berens, S.C.、Horner, A.J.、Franklin, A.: 大脑中颜色的分类编码。美国国家科学院院刊 111(12), 4590–4595(2014)4。Brock, A., Donahue, J., Simonyan, K.: 用于高保真自然图像合成的大规模 gan 训练。arXiv 预印本 arXiv: 1809.11096 (2018)5。Brock, A., Donahue, J., Simonyan, K.: 用于高保真自然图像合成的大规模 GAN 训练。见: 第七届学习表征国际会议, ICLR 2019, 美国洛杉矶新奥尔良, 2019 年 5 月 6-9 日。OpenReview.net (2019)6。Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A.等人: 语言模型是很少学习的。神经信息处理系统的进展 33, 1877–1901 (2020)7。Chen, Z., Qing, J., Xiang, T., Yue, W.L., Zhou, J.H.: 超越大脑: 用于视觉解码的稀疏掩模建模条件扩散模型。arXiv 预印本 arXiv: 2211.06956 (2022)8。陈志、清杰、向天、岳文林、周建华: 超越大脑: 人类视觉解码的掩模建模条件扩散模型。见: arXiv (2022 年 11 月), <https://arxiv.org/abs/2211.06956>。Davis, K.M.、de la Torre-Ortiz, C., Ruotsalo, T.: 大脑监督图像编辑。见: IEEE/CVF 计算机视觉和模式识别会议论文集。第 18480–18489 页 (2022)10。Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: 用于语言理解的深度双向变换器的预训练。arXiv 预印本 arXiv: 1810.04805(2018)11。Dhariwal, P., Nichol, A.: 扩散模型在图像合成方面击败了甘斯。神经信息处理系统进展 34, 8780–8794 (2021)12。丁明、郑文、洪文、唐杰: Cogview2: 通过分层转换器更快更好地生成文本到图像。arXiv 预印本 arXiv: 2204.14217(2022)13。Doersch, C., Gupta, A., Efros, A.A.: 通过上下文预测进行无监督视觉表示学习。见: IEEE 计算机视觉国际会议论文集。第 1422–1430 页 (2015)14。Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., 等人: 一张图像相当于 16x16 个单词: 用于大规模图像识别的 Transformers。arXiv preprint arXiv:2010.11929 (2020)15。Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: 生成对抗网络。ACM 通讯 63(11), 139–144 (2020)16。Hadsell, R., Chopra, S., LeCun, Y.: 通过学习不变映射来降维。见: 2006 年 IEEE 计算机学会计算机视觉和模式识别会议 (CVPR'06)。卷。2, 第 1735–1742 页。IEEE (2006) 17。He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: 蒙面自动编码器是可扩展的视觉学习器。见: IEEE/CVF 计算机视觉和模式识别会议论文集。第 16000–16009 页 (2022)

18. Ho, J., Jain, A., Abbeel, P.: 去噪扩散概率模型。神经信息处理系统进展 33, 6840–6851 (2020)19。Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: 用于高保真图像生成的级联扩散模型。J.马赫.学习。资源。23, 47:1–47:33 (2022)20。Jayaram, V., Barachant, A.: Moabb: bcis 的值得信赖的算法基准测试。神经工程学杂志 15(6), 066011 (2018)21。Karras, T., Laine, S., Aila, T.: 用于生成对抗网络的基于样式的生成器架构。见: IEEE/CVF 计算机视觉和模式识别会议论文集。第 4401–4410 页 (2019)22。Kavasidis, I., Palazzo, S., Spampinato, C., Giordano, D., Shah, M.: Brain2Image: 将大脑信号转换为图像。载于: 2017 年 ACM 多媒体会议记录, MM 2017, 美国加利福尼亚州山景城, 2017 年 10 月 23-27 日。第 1809-1817 页。ACM (2017) 23。Kong, Z., Ping, W.: 扩散概率模型的快速采样。CoRRabs/2106.00132 (2021) , <https://arxiv.org/abs/2106.00132>。Mozafari, M., Reddy, L., VanRullen, R.: 使用 bigbigan 从 fmripattern 重建自然场景。见: 2020 年国际神经网络联合会议 (IJCNN)。第 1-8 页。IEEE (2020) 25。Noroozi, M., Favaro, P.: 通过解决拼图游戏进行视觉表征的无监督学习。见: 计算机视觉 - ECCV 2016: 第 14 届欧洲会议, 荷兰阿姆斯特丹, 2016 年 10 月 11-14 日, 会议记录, 第六部分。第 69–84 页。Springer (2016)26。Ozcelik, F., Choksi, B., Mozafari, M., Reddy, L., VanRullen, R.: 使用实例条件甘斯从 fmri 模式和语义大脑探索中重建感知图像。见: 2022 年国际神经网络联合会议 (IJCNN)。第 1-8 页。IEEE (2022) 27。Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. 等al.: 从自然语言监督中学习可转移的视觉模型。见: 机器学习国际会议。第 8748–8763 页。PMLR (2021)28。Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. 等人: 语言模型是无监督的多任务学习者。OpenAI 博客 1(8), 9 (2019)29。Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: 具有剪辑潜伏的分层文本条件图像生成。arXiv 预印本 arXiv: 2204.06125(2022)30。Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: 零样本文本到图像生成。见: 国际机器学习会议。第 8821–8831 页。PMLR (2021)31。Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 使用潜在扩散模型进行高分辨率图像合成。见: IEEE/CVF 计算机视觉和模式识别会议论文集。第 10684–10695 页 (2022)32。Ronneberger, O., Fischer, P., Brox, T.: U-net: 用于生物医学图像分割的卷积网络。见: 医学图像计算和计算机辅助干预 - MICCAI 2015: 第 18 届国际会议, 德国慕尼黑, 2015 年 10 月 5-9 日, 会议记录, 第 III 部分 18。第 234-241 页。施普林格 (2015) 33。Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour,S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G. 等人: 具有深入语言理解的真实感文本到图像扩散模型。arXiv 预印本 arXiv:2205.11487 (2022)

34. San-Roman, R., Nachmani, E., Wolf, L.: 生成扩散模型的噪声估计。CoRR 绝对/2104.02600 (2021)35. Shen, G., Dwivedi, K., Majima, K., Horikawa, T., Kamitani, Y.: 根据人脑活动进行端到端深度图像重建。前沿计算。神经科学。13,21 (2019) 36. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: 使用非平衡热力学的深度无监督学习。见：国际机器学习会议。第 2256–2265 页。PMLR (2015)37. Song, J., Meng, C., Ermon, S.: 去噪扩散隐式模型。见：第九届学习表征国际会议，ICLR 2021，虚拟活动，奥地利，2021 年 5 月 3-7 日。OpenReview.net (2021)38. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: 通过随机微分方程基于分数的生成模型。见：第九届学习表征国际会议，ICLR 2021，虚拟活动，奥地利，2021 年 5 月 3-7 日。OpenReview.net (2021)39. Takagi, Y., Nishimoto, S.: 利用人脑活动的潜在扩散模型进行高分辨率图像重建。bioRxiv, 第 2022-11 页 (2022)40. Tirupattur, P., Rawat, Y.S., Spampinato, C., Shah, M.: Thoughtviz: 使用生成对抗网络可视化人类思想。见：2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, 韩国首尔, 2018 年 10 月 22-26 日。第 950–958 页。ACM (2018) 41. Vahdat, A., Kreis, K., Kautz, J.: 潜在空间中基于分数的生成模型。见：Ranzato, M., Beygelzimer, A., Dauphin, Y.N., Liang, P., Vaughan, J.W. (编辑) 神经信息处理系统的进展 34: 2021 年神经信息处理系统年会, NeurIPS 2021, 2021 年 12 月 6-14 日, 虚拟。第 11287–11302 页 (2021)42. Wang, X., Gupta, A.: 使用视频进行视觉表示的无监督学习。见：IEEE 计算机视觉国际会议论文集。第 2794–2802 页 (2015)43. 谢Z., 张Z., 曹Y., 林Y., 鲍J., 姚Z., 戴Q., 胡H.: Simmim: 掩模图像建模的简单框架。见：IEEE/CVF 计算机视觉和模式识别会议论文集。第 9653–9663 页 (2022)

Conference on Computer Vision and Pattern Recognition. pp. 9653–9663 (2022)