arXiv:2301.08562v1 [cs.LG] 9 Jan 2023

潜在自回归源分离

Emilian Postolache[*1]、**Giorgio Mariani**[*1]、**Michele Mancusi**[*1]、**Andrea Santilli**[1],**Luca Cosmo**[†2]、**Emanuele Rodol`a**[†1]

[1] 罗马智慧大学，意大利 [2] 威尼斯卡福斯卡里大学，意大利

postolache@di.uniroma1.it、mariani@di.uniroma1.it、mancusi@di.uniroma1.it

摘要

自回归模型在广泛的领域中都取得了令人瞩目的成果，无论是生成质量还是下游任务表现。在连续域中，这些成果背后的关键因素之一是使用量化潜在空间（例如通过VQ-VAE自动编码器获得），这有助于降维并缩短推理时间。然而，使用现有的预训练模型执行新的非平凡任务并不容易，因为这需要额外微调或大量训练才能通过提示引导模型。本文提出*LASS*，以便在无需额外的潜在自回归源分离 基于梯度的优化或修改现有模型的情况下，实现向量量化分离（即将输入信号解混为其组成源）。我们提出的分离方法基于贝叶斯形式，其中自回归模型充当先验，并通过统计加数词元潜在和的频数，构造离散（非参数）似然函数。我们在图像和音频上测试了该方法，并采用了多种采样策略（例如祖先采样、束搜索）。结果表明，该方法的分离质量与现有方法不相上下，同时显著缩短了推理时间，并提高了对更高维数据的可扩展性。

引言

自回归模型在众多领域都取得了令人瞩目的成果，涵盖自然语言（Brown 等，2020）以及音频（Dhariwal 等，2020）和视觉（Razavi、van den Oord 和 Vinyals，2019；Esser、Rombach 和 Ommer，2021）等密集值领域，还包括多模态联合空间（Ramesh 等，2021；Yu 等，2022）。在密集场景中，通常会训练自回归模型，使其处理通过连续数据量化获得的离散潜在表示，也可能使用 VQ-VAE 自动编码器（van den Oord、Vinyals 和 Kavukcuoglu，2017）。这样一来，便能在缩短推理时间的同时生成更高分辨率的样本。此外，学习到的潜在表示也有助于下游任务（Castellon、Donahue 和 Liang，2021）。然而，要执行新的、非平凡的任务，标准做法是微调模型，或者通过扩大训练规模来引导模型响应提示（Wei 等，2021；Sanh 等，2022）。前一种方式通常是默认选择，但需要额外的优化步骤或对模型进行修改。后一种方式在非平凡任务上颇具挑战，尤其是在不同于自然语言的领域中（Yang 等，2022；Hertz 等，2022）。

本文旨在解决此类任务之一，即源分离，并利用现有的矢量量化自回归模型，无需进行任何基于梯度的优化或架构修改。随着深度学习取得成功，从混合信号中分离两个或更多源的任务近期受到了广泛关注，尤其是在音频领域，涵盖语音（Dovrat、Nachmani 和 Wolf，2021）、音乐（D´efossez，2021）以及通用源分离（Wisdom 等，2021；Postolache 等，2022）。尽管图像源分离不像音频领域的对应任务那样广为人知，相关研究也已对此展开探讨（Halperin、Ephrat 和 Hoshen，2019）。大多数成功方法通过明确的监督取得显著成果（Luo 和 Mesgarani，2019；D´efossez 等，2019），或利用大规模无监督回归（Wisdom 等，2020）。

我们提出一种生成式方法，通过在潜在 VQ-VAE 域上训练自回归先验分布来执行源分离（使用类别信息时，该方法为弱监督；否则为无监督）。通过统计潜在混合词元相对于源词元的出现次数，学习一个非参数稀疏似然函数；这些词元是将数据域求和信号及其相对加数经由 VQ-VAE 映射得到的。该模块既不会侵入 VQ-VAE，也不会侵入自回归先验，因为在学习似然函数的过程中，VQ-VAE 的表示空间不会发生变化。最后，在推理时间，通过贝叶斯公式将似然函数与自回归先验的估计相结合，从而得到后验分布。通过标准离散采样器（例如，祖先采样、束搜索）从后验分布中获得分离结果。我们将该方法称为 *LASS*（潜在自回归源分离）。

我们的贡献总结如下：

- 我们提出*LASS*，这是一种用于源分离的贝叶斯推断方法，能够利用现有的、在量化潜在域中预训练的自回归模型。
- 我们在图像域中对*LASS* 进行了实验，并

[*]贡献相同，作者排序随机。

[†]共同担任末位作者。

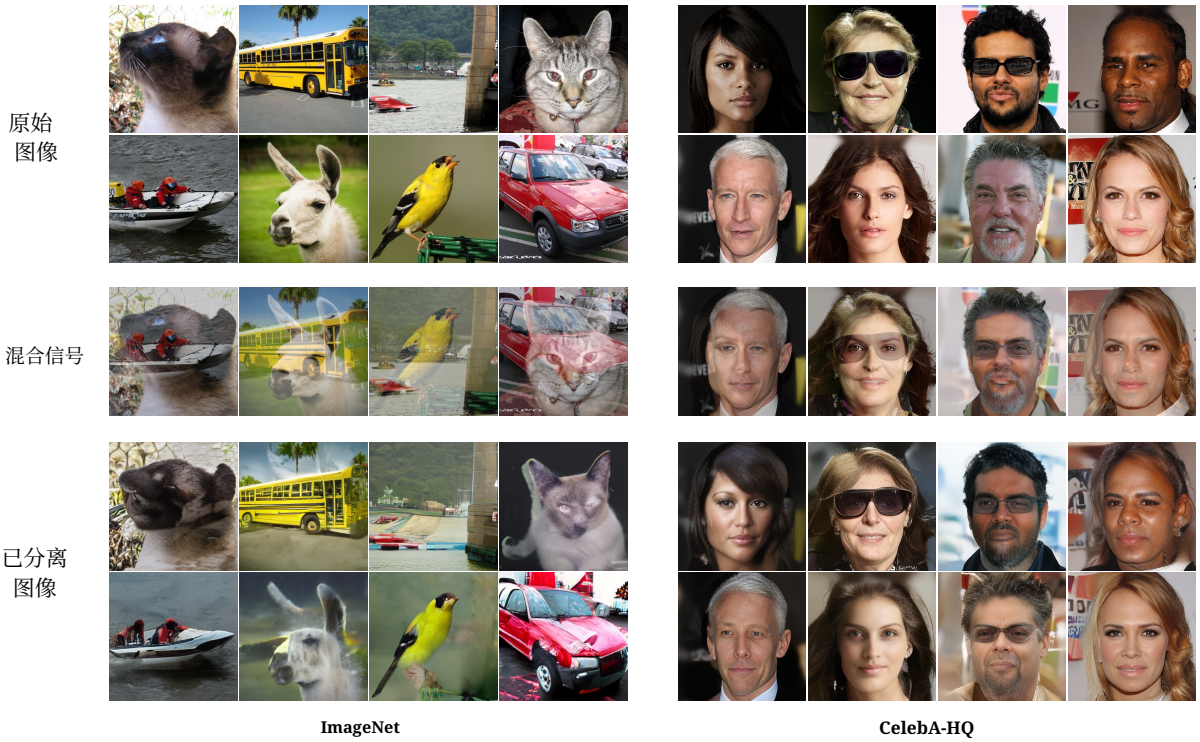

图1：使用预训练自回归模型，通过 *LASS* 获得的 256x256 分离结果。左：类别条件 ImageNet。右：无条件 CelebA-HQ。

展示了在 MNIST 和 CelebA 上，相比竞争方法以显著更低的推理时间成本取得的竞争力结果（32×32）。得益于*LASS* 对预训练模型的可扩展性，我们还展示了在 ImageNet （256×256）和 CelebA-HQ （256×256）上的定性结果。据我们所知，这是首个将生成式源分离扩展到更高分辨率图像的方法。

- 我们在 Slakh2100 数据集上针对音乐源分离任务开展了 *LASS* 实验。与生成式竞争方法相比，*LASS* 取得了与最先进的监督模型相当的性能，同时显著降低了推理和训练时间成本。

相关工作

源分离问题传统上以无监督方式处理，统称为盲源分离（Comon，1994；Hyv¨arinen 和 Oja，2000；Huang 等，2012；Smaragdis 等，2014）。在这种设定下，没有关于要从混合信号中分离出的源的信息。因此，这些方法依赖一些宽泛的数学先验，例如源独立性（Hyv¨arinen 和 Oja，2000）或重复性（Rafii 和 Pardo，2012）来实现分离。随着深度学习的兴起，大多数主要的源分离方法可归类为基于回归的方法或基于生成的方法。

基于回归的源分离

在这种设定下，混合信号被输入参数化模型（即神经网络），模型输出分离后的源。训练通常以监督方式进行，通过回归损失（例如 $\mathcal{L}_1$ 或 $\mathcal{L}_2$）使估计的分离结果与真实源相匹配（Gus´o 等，2022）。监督回归已应用于图像源分离（Halperin、Ephrat 和 Hoshen，2019），但主要是在音频领域得到研究，其中有两种常见方法：基于掩码的方法和波形方法。在基于掩码的方法中，模型通过将估计掩码应用于混合信号来执行分离，通常是在短时傅里叶变换域（Roweis，2000；Uhlich、Giron 和 Mitsufuji，2015；Huang 等，2014；Nugraha、Liutkus 和 Vincent，2016；Liu 和 Yang，2018；Takahashi、Goswami 和 Mitsufuji，2018）。在波形方法中，模型直接在时域输出估计的源，从而避免相位估计；将信号从短时傅里叶变换域转换到波形域时需要进行相位估计（Llu´ís、Pons 和 Serra，2019；Luo 和 Mesgarani，2019；D´efossez 等，2019）。

生成式源分离

随着深度生成模型取得成功（Goodfellow 等，2014；Kingma 和 Welling，2014；Ho、Jain 和 Abbeel，2020；Song 等，2021），一类新的生成式源分离方法正日益受到关注。这种设定强调利用广泛的生成

模型（尤其是预训练模型）来解决分离任务，而无需专用架构（如基于回归的模型所采用的架构）。

继基于生成对抗网络的深度生成式分离早期工作之后（Subakan 和 Smaragdis 2018；Kong 等人 2019；Narayanaswamy 等人 2020），Jayaram 和 Thickstun（2020）提出了图像场景下使用基于分数的模型（Song 和 Ermon 2019）的生成式分离方法 BASIS（BASIS-NCSN），以及基于流模型并经噪声退火的版本（BASIS-Glow）。该推理过程通过图像域中的朗之万动力学（Parisi 1981）进行，并取得了良好的定量和定性结果。作者通过采用噪声调度重新训练自回归模型，将朗之万动力学推理过程扩展到自回归模型，并提出了并行灵活（PnF）方法（Jayaram 和 Thickstun 2021）。尽管这一方法颇具创新性，尤其适用于图像修复等任务，但它无法直接使用预训练自回归模型，需要在不同噪声级别下进行微调。此外，由于直接在数据域中运行，该方法推理时间较长，也难以扩展到更高分辨率。本文通过提出一种针对潜在自回归模型的分离过程，延续了这一研究方向；该过程无需重新训练，可扩展至任意预训练检查点，并兼容标准离散采样器。

背景

本节简要介绍矢量量化自编码器（VQ-VAE）和自回归模型，因为它们是*LASS*所采用的分离过程的核心组成部分。

VQ-VAE

数据点$\mathbf{x} \in \mathbb{R}^N$（$N$ 表示数据点的总长度，例如音频序列的长度或图像中的像素通道数）可以通过 VQ-VAE（van den Oord、Vinyals 和 Kavukcuoglu，2017年）映射到离散潜在域。首先，编码器 $E_\theta: \mathbb{R}^N \to \mathbb{R}^{S\times C}$将$\mathbf{x}$映射为 $E_\theta(\mathbf{x}) = (\mathbf{h}_1, \dots, \mathbf{h}_S)$，其中 C表示潜在通道数，S 表示潜在序列的长度。瓶颈块 $B: \mathbb{R}^{S\times C} \to [K]^S$ 将编码转换为离散序列$\mathbf{z} = (z_1, \dots, z_S)$，方法是将每个 $\mathbf{h}_s$ 映射到学习得到的向量的（有序）集合$\mathcal{C} = \{\mathbf{e}_k\}_{k=1}^K$（称为编码）中最近邻$\mathbf{e}_{z_s}$ 的索引（也称为词元）$z_s = B(\mathbf{h}_s)$。解码器 $D_\psi: [K]^S \to \mathbb{R}^N$将潜在序列映射回数据域，得到重建 $\hat{\mathbf{x}} = D_\psi(\mathbf{z})$。VQ-GAN（Esser、Rombach 和 Ommer，2021年）是 VQ-VAE 的增强版本，其训练损失加入了判别器和感知损失，从而在提高自编码器压缩率的同时改善重建质量。关于 VQ-VAE 和 VQ-GAN 的更多详情，请参阅（van den Oord、Vinyals 和 Kavukcuoglu，2017年）和（Esser、Rombach 和 Ommer，2021年）。在本文余下部分中，我们将两种模型统称为 VQ-VAE，并在必要时加以区分。

自回归模型

自回归模型定义了离散域 $[K]^S$（在我们的例子中，是 VQ-VAE 的潜在域）上的概率分布。序列$\mathbf{z} = (z_1, \dots, z_S)$的联合概率可通过链式法则分解为

$$p_\phi(\mathbf{z}) = \prod_{s=1}^{S} p_\phi(z_s|\mathbf{z}_{<s}), \tag{1}$$

其中，$p_\phi(\cdot)$是一个学习得到的参数化模型，通常是神经网络，例如卷积神经网络（van den Oord 等，2016；Salimans 等，2017）或 Transformer（Vaswani 等，2017）。在推理时，样本可根据所选的采样过程生成。通常采用祖先采样，即在每一步中，从条件分布 $p_\phi(z_s|\mathbf{z}_{<s})$中随机抽取词元 z_s；也可以采用 top-k 筛选（Kool、van Hoof 和 Welling，2020），以提高生成数据的多样性（Holtzman 等，2020）。如果目标是最大化整个序列的概率（相对于所有序列），则会使用束搜索等启发式方法（Reddy 等，1977）。束搜索会在推理过程中并行保留 B 个可能的假设（波束）$\mathbf{z}^1, \dots, \mathbf{z}^B$。在每一步 s，它都会为每个波束计算条件分布 $p_\phi(z_s^b|\mathbf{z}_{<s}^b)$，并选出 B 个能使联合分布 $p_\phi(\mathbf{z}_{<s}^b)p_\phi(z_s|\mathbf{z}_{<s}^b)$最大的假设。

方法

设 $\mathbf{x} = (\mathbf{x}^1, \mathbf{x}^2) \in \mathbb{R}^{2\times N}$ 表示两个服从 $p_{\text{data}} = (p^1_{\text{数据}}\ p^2_{\text{数据}})$ 分布的源，而 $\mathbf{y} = (\mathbf{x}^1 + \mathbf{x}^2)/2$ 表示可观测混合信号。生成式源分离的目标是利用贝叶斯后验（假设各源相互独立），根据混合信号估计源$\mathbf{x}$：

$$p(\mathbf{x}^1, \mathbf{x}^2|\mathbf{y}) \propto p^1_{\text{data}}(\mathbf{x}^1)p^2_{\text{data}}(\mathbf{x}^2)p(\mathbf{y}|\mathbf{x}^1, \mathbf{x}^2). \tag{2}$$

直接在连续数据域中使用式（2）效率低下。为了解决这一问题，我们首先在 VQ-VAE 的潜在空间中使用自回归模型对 p_{data}进行建模。通过改变数据域，我们接着重新定义似然函数 $p(\mathbf{y}|\mathbf{x}^1, \mathbf{x}^2)$，从而无需基于梯度的优化或模型重新训练。我们将在下一小节讨论第一个问题，并在再下一小节讨论第二个问题。然后，我们将介绍如何使用 *LASS* 进行推理以分离数据，并提出一种推理后细化过程。

潜在自回归源分离

本文探讨了以下情形：所有源都由同一个自回归模型 p_ϕ 估计（无监督[1]）；以及每个源分别由两个独立模型 $p_\phi = (p_{\phi_1}, p_{\phi_2})$ 估计（弱监督），这两个模型可以按类别条件训练，也可以独立训练。下文将重点讨论后一种情形，因为前一种情形可以通过设定$p_{\phi_1} = p_{\phi_2}$推广得到。

[1]请勿与无监督盲设定混淆；也就是说，在无监督设定中，我们可以访问源，但没有类别标签。

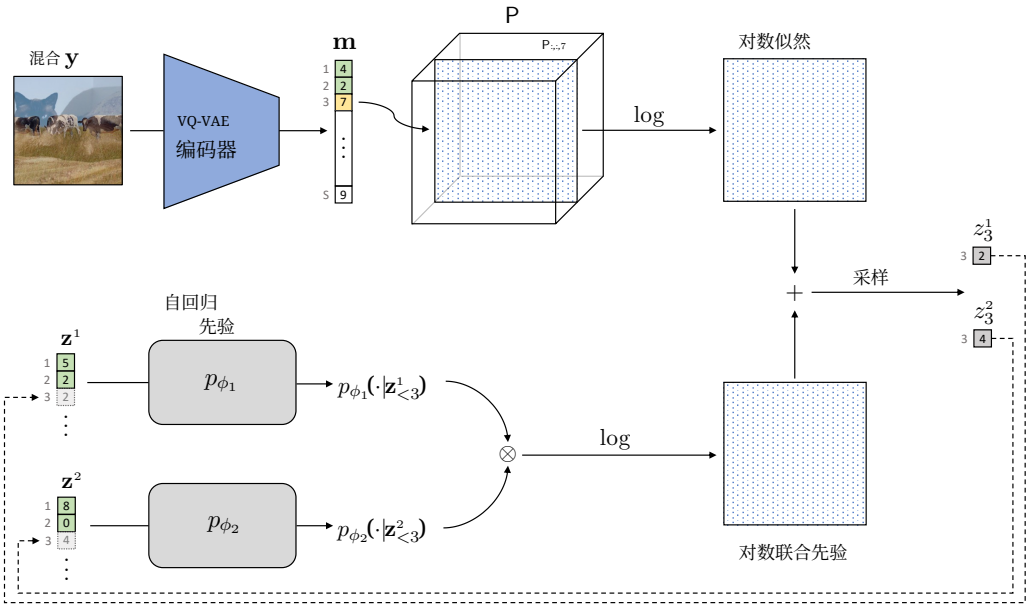

图2：*LASS* 分离过程示意图。该图展示了 $s = 3$时的分离过程，并重复执行直至 $s = S$。推理结束时，我们通过 VQ-VAE 解码器（图中未显示）对$\mathbf{x}^1$ 和$\mathbf{x}^2$ 进行解码，得到$\mathbf{z}^1$ 和$\mathbf{z}^2$ 。更多详情请参阅算法1。

我们分别用 $\mathbf{z} = (\mathbf{z}^1, \mathbf{z}^2) = B(E_\theta(\mathbf{x}))$ 和 $\mathbf{m} = B(E_\theta(\mathbf{y}))$ 表示潜在源和混合信号。式（2）中的后验分布可在潜在域中局部表示为

$$p(\mathbf{z}_s|\mathbf{z}_{<s}, \mathbf{m}_{\leq s}) \propto p_\phi(\mathbf{z}_s|\mathbf{z}_{<s}) p(\mathbf{m}_{\leq s}|\mathbf{z}_{\leq s}), \quad (3)$$

对于所有 $s = 1$、.....、S。第一个因子是（联合）贝叶斯先验，由自回归分布建模。第二个因子是似然函数，用于量化序列 $\mathbf{z}^1_{\leq s}$、$\mathbf{z}^2_{\leq s}$ 组合成 $\mathbf{m}_{\leq s}$ 的可能性。

由于卷积向量量化变分自编码器中的每个编码都描述数据的一个局部区域，并且数据域中的混合操作是逐点进行的，因此潜在编码之间的混合关系在潜在域中也是局部的。因此，我们可以在式（3）的似然函数中去掉对先前上下文的依赖，将其近似为

$$p(\mathbf{m}_{\leq s}|\mathbf{z}_{\leq s}) \approx p(m_s|\mathbf{z}_s). \quad (4)$$

注意，由于不依赖全局上下文，也就不依赖序列中的特定位置，因此我们可以去掉位置索引 s:

$$p(m_s|\mathbf{z}_s) = p(m_s|z^1_s, z^2_s) = p(m|z^1, z^2). \quad (5)$$

下一小节将介绍 *LASS* 如何对似然函数建模。

用于源分离的离散似然函数

以往的生成式源分离研究（Jayaram 和 Thickstun 2020年、2021年）直接在数据域中对似然函数建模，通常采用一个 σ-各向同性高斯项

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|(\mathbf{x}^1 + \mathbf{x}^2)/2, \sigma^2\mathbf{I}). \quad (6)$$

在我们的设定中，由于 VQ-VAE 并未对潜在空间施加明确的算术结构，因此我们无法通过规范的求和运算组合 z^1_s 和 z^2_s（或对应的稠密编码 $\mathbf{e}^{z^1_s}$ 和 $\mathbf{e}^{z^2_s}$）。

为了解决这一问题，我们使用离散条件分布对式（5）中的似然函数进行建模，并用三阶张量[2] $\mathsf{P} \in \mathbb{R}^{K\times K\times K}$表示:

$$p(\cdot\,|z^1, z^2) = \mathsf{P}_{z^1, z^2, :} \quad (7)$$

为了学习 P，我们遍历数据集 X，根据潜在源的词元，对潜在混合词元进行频数统计。我们首先将空整数张量 $\mathsf{F}^0 \in \mathbb{N}^{K\times K\times K}$ 初始化。遍历 $\mathbf{x}^1, \mathbf{x}^2 \in X$，计算 $\mathbf{y} = (\mathbf{x}^1 + \mathbf{x}^2)/2$，然后得到潜在序列$\mathbf{z}^1 = B(E_\theta(\mathbf{x}^1)), \mathbf{z}^2 = B(E_\theta(\mathbf{x}^2))$ 和 $\mathbf{m} = B(E_\theta(\mathbf{y}))$。对于步骤 t 中的每个条目 $(z^1_s, z^2_s, m_s) \in (\mathbf{z}^1, \mathbf{z}^2, \mathbf{m})$，我们只需将之前的计数加一:

$$\mathsf{F}^t_{z^1_s, z^2_s, m_s} = \mathsf{F}^{t-1}_{z^1_s, z^2_s, m_s} + 1\,, \quad (8)$$

$$\mathsf{F}^t_{z^2_s, z^1_s, m_s} = \mathsf{F}^{t-1}_{z^2_s, z^1_s, m_s} + 1\,. \quad (9)$$

我们对加数的顺序进行置换，以确保求和具有交换性。完成统计后，我们可以将 P 定义为:

$$\mathsf{P}_{z^1, z^2, :} = \frac{1}{\sum_{m=1}^{K} \mathsf{F}_{z^1, z^2, m}} \mathsf{F}_{z^1, z^2, :} \quad (10)$$

在推理时，似然函数（以 z^1 和 z^2 为参数，并固定 m）可通过对张量进行切片获得

[2]我们遵循 Goodfellow、Bengio 和 Courville（2016）中关于张量的记法。

算法1：*LASS* 推理

输入： $\mathbf{y}$

输出： $\mathbf{x}^1$,$\mathbf{x}^2$

1: **米** $\leftarrow B(E_\theta(\mathbf{y}))$
2: $\mathbf{z}^1 \leftarrow []$
3: $\mathbf{z}^2 \leftarrow []$
4: **对** $s = 1$至 S **执行**
5: 先验 $\leftarrow \log(p_{\phi_1}(\cdot\,|\mathbf{z}^1) \otimes p_{\phi_2}(\cdot\,|\mathbf{z}^2))$
6: 似然 $\leftarrow \log(\mathsf{P}_{:,:,m_s})$
7: 后验分布 $\leftarrow$ 先验 $+\,\lambda$ 似然
8: $(z_s^1,\ z_s^2) \leftarrow$ 采样器(后验分布)
9: $\mathbf{z}^1 \leftarrow$ 拼接$(\mathbf{z}^1, z_s^1)$
10: $\mathbf{z}^2 \leftarrow$ 连接$(\mathbf{z}^2,\ z_s^2)$
11: **结束循环**
12: $\mathbf{x}^1 \leftarrow D_\psi(\mathbf{z}^1)$
13: $\mathbf{x}^2 \leftarrow D_\psi(\mathbf{z}^2)$
14: **返回** $\mathbf{x}^1$, $\mathbf{x}^2$

沿 m，即

$$p(m|\cdot,\cdot) = \mathsf{P}_{:,:,m}. \tag{11}$$

乍看之下，不带参数地对条件分布进行建模似乎会造成内存效率低下，其复杂度为 $O(K^3)$。实际上，张量 P 高度稀疏。我们在表1中展示了所有实验的这一特性，其中 P 的密度定义为 P 中非零元素所占的百分比。

在 VQ-VAE 的潜在域中采用离散似然函数进行源分离是一种灵活的方法；无需更改 VQ-VAE 的表示，非参数学习过程不依赖超参数，而且自回归先验无需重新训练。

推理过程

给定可观测混合信号 $\mathbf{y}$、自回归先验p_{ϕ_1}，p_{ϕ_2} 以及学习得到的似然张量 P，可以按照算法（1）所述并如图 2 所示执行推理，并估计 $\mathbf{x}^1$,$\mathbf{x}^2$,。

首先将 $\mathbf{y}$ 映射到潜在域，得到$\mathbf{m} = B(E_\theta(\mathbf{y}))$，并用空序列初始化估计值 $\mathbf{z}^1$,$\mathbf{z}^2$。算法对 $s = 1, \ldots, S$进行迭代。

每一步中，联合先验（一个 $K \times K$ 矩阵）通过对自回归模型根据过去上下文预测的两个分布取外积来计算（第 5 行）。为保证数值稳定性，我们使用这些分布的对数。接着计算对数似然函数（第 6 行），即对 $\mathsf{P}_{:,:,m_s}$取对数。在我们的实验中，可以对对数似然应用不同的缩放因子 λ，使其与先验相平衡。随后将两个矩阵合并，在第 7 行得到后验分布。

最后（第 8–10 行），可以采用不同技术从后验分布中采样出最佳候选词元（z_s^1，z_s^2）。在我们的实验中，我们使用了祖先采样（采用和不采用 top-k 筛选）以及束搜索。推理循环结束后，估计序列通过 VQ-VAE 的解码器映射回数据域（第 12–13 行），得到$\mathbf{x}^1$ 和$\mathbf{x}^2$。

数据集	K	密度（%）
MNIST	256	1.49×10^{0}
CelebA	512	6.06×10^{0}
CelebA-HQ	1024	3.80×10^{-1}
ImageNet	16384	3.90×10^{-3}
SLAKH（鼓 + 贝斯）	2048	7.60×10^{-2}

表1：不同数据集上似然函数的统计信息。 K 是 VQ-VAE（或 VQ-GAN）潜在编码的数量。密度是似然函数中非零元素所占的百分比。

推理后细化 分离图像的质量受 VQ-VAE 自动编码器生成图像的质量限制。为了提升分离结果，我们可以采用额外的细化步骤，迭代优化样本的 VQ-VAE 潜在表示：

$$\mathbf{e}_{t+1}^1 = \mathbf{e}_t^1 + \alpha \nabla_{\mathbf{e}_t^1} \|D_\psi(\mathbf{e}_t^1) + D_\psi(\mathbf{e}_t^2) - 2\mathbf{y}\|_2 \tag{12}$$

$$\mathbf{e}_{t+1}^2 = \mathbf{e}_t^2 + \alpha \nabla_{\mathbf{e}_t^2} \|D_\psi(\mathbf{e}_t^1) + D_\psi(\mathbf{e}_t^2) - 2\mathbf{y}\|_2 \tag{13}$$

对于 $t = 1, \ldots, T-1$ 和 $\mathbf{e}_1^1 = E_\theta(\mathbf{x}^1), \mathbf{e}_1^2 = E_\theta(\mathbf{x}^2)$。简单来说，我们优化稠密的潜在嵌入，使其解码结果相加后能更好地还原混合信号，并将其初始化为算法1的输出。在 MNIST 数据集上，这种策略尤其有效；我们通过逐像素指标（峰值信噪比）评估分离质量，而 VQ-VAE 往往会生成平滑的图像。

实验

我们在各种数据集上进行了定量和定性实验，以展示*LASS*的有效性和可扩展性。在图像域中，我们在 MNIST（Lecun 等，1998）和 CelebA（32×32）（Liu 等，2015）上进行评估，并展示更高分辨率数据集 CelebA-HQ（256×256）（Karras 等，2018）和 ImageNet（256×256）（Deng 等，2009）上的定性结果。在音频领域，我们在 Slakh2100（Manilow 等，2019）上进行测试，这是一个适用于生成建模的大型音乐源分离数据集。所有实验均在配备 24 GB 显存的单张 Nvidia RTX 3090 GPU 上完成。所有模型的实现细节均列于配套网站[3]。

图像源分离

我们选择 Transformer架构（Vaswani 等，2017年）作为所有图像源分离实验的自回归骨干网络。对于 MNIST 和 CelebA，我们首先训练一个 VQ-VAE，然后在其潜在空间中训练自回归 Transformer。考虑到 CelebA 的变化更多、重建数据所需的信息也更多，我们在 MNIST 上使用 $K = 256$ 编码，在 CelebA 上使用 $K = 512$。对于 CelebA-HQ 和 ImageNet，我们采用预训练VQ-GAN（Esser、Rombach 和 Ommer，2021）以及作者发布的预训练 Transformer[4]

[3]github.com/gladia-research-group/latent-autoregressive-source-separation[4]github.com/CompVis/taming-transformers

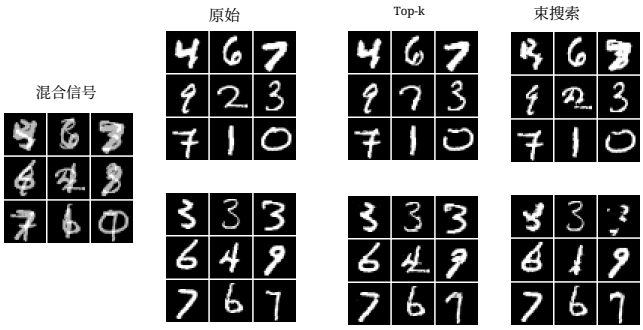

图3：对随机批次样本使用 top-k 采样（$k = 32$）在 MNIST 上的结果。Top-k 采样生成的数字轮廓更清晰，与表3中的结果一致。

分离方法	MNIST（峰值信噪比）	CelebA（弗雷歇 inception 距离）
平均值	14.9	15.19
NMF	9.4	-
S-D	18.5	-
BASIS Glow	22.7	-
BASIS NCSN	29.3	7.55
***LASS*（我们的方法）**	24.2	8.96

表 2：在 MNIST 和 CelebA 测试集上与其他方法的比较。结果以峰值信噪比（越高越好）和弗雷歇 inception 距离（越低越好）报告。

（用于 CelebA-HQ 的 celebahq Transformer 检查点，以及用于 ImageNet 的 cin Transformer）。鉴于 *LASS* 的灵活性，我们直接将其用于分离算法，不作任何修改。在 CelebA-HQ 上，VQ-GAN 有 $K = 1024$ 个编码，而在 ImageNet 上有 $K = 16384$个。首先，在所有基于图像的实验中，我们都使用“方法”一节中介绍的过程来学习 P张量。如表1所示，CelebA 的稀疏度最低（密度最高），而 ImageNet 的稀疏度最高。所有情况下，密度都低于 7%，且推理过程不会受到内存问题的影响。

定量结果 为了评估*LASS*生成的图像分离质量，我们将本方法与 MNIST 和 CelebA 上的不同基线方法进行比较。

在 MNIST 上，我们将*LASS* 与以下方法的报告结果进行比较：两种生成式分离方法“BASIS NCSN”（基于分数）和“BASIS Glow”（基于噪声退火流），来自(Jayaram 和 Thickstun2020)；基于 GAN 的“S-D”方法（Kong 等，2019）；完全监督版本的 Neural Egg“NES”；以及“平均值”基线，其中分离结果直接从混合信号$\mathbf{x}^1 = \mathbf{x}^2 = \mathbf{y}/2$中获得。在所有这些情况下，评估指标均为 PSNR（峰值信噪比）（Hor´e 和 Ziou 2010）。我们遵循 (Jayaram 和 Thickstun 2020) 在 MNIST 上的实验流程，对由 12,000 个测试源组合而成的 6,000 个混合信号进行分离。为了为该数据集选出最佳采样器，我们在由测试集划分构造的 1,000 个混合信号上验证了表3中的采样器集合。我们发现，随机

采样方法	MNIST（峰值信噪比）	SLAKH（信号失真比）
贪心	17.36 ± 5.90	1.23 ± 2.33
束搜索	16.96 ± 5.78	5.01 ± 2.39
祖先采样	24.03 ± 6.37	4.23 ± 2.29
前-k ($k = 16$)	23.74 ± 6.55	3.13 ± 2.53
前-k ($k = 32$)	24.23 ± 6.23	2.93 ± 2.20
前k （$k = 64$）	23.85 ± 6.13	3.24 ± 3.29

表3：*LASS* 采用不同采样方法时的性能。在 MNIST 上，报告的分数为 PSNR（dB）（越高越好）；在 SLAKH 上，报告的分数为 SDR（dB）（越高越好）。使用随机采样器（祖先采样或 top-k）时，批次中选出的解是其总和与输入混合信号之间的 $\mathcal{L}_2$ 距离最小的解。

采样器表现最佳（峰值信噪比为 > 20 dB），而 MAP 方法未能达到令人满意的性能。我们推测，在 MNIST 数字这类稀疏图像上进行早期推理时，束搜索容易因选择错误而陷入次优解。采用 $k = 32$ 的前k 采样效果最佳，因此我们选择该方法进行评估（定性比较见图3）。对于测试集中的每个混合，我们采样得到一批包含 512 个分离结果的候选项，选出各分离结果之和与混合最匹配的结果（依据 $\mathcal{L}_2$ 距离），最后使用 $T = 500$ 和 $\alpha = 0.1$ 执行公式（12）、（13）中的细化过程。该实验的评估指标见表2，推理时间见表4。得益于潜变量量化，我们的方法取得了优于“NMF”、“S-D”和“BASIS Glow”的指标，且速度快于“BASIS NCSN”。后者取得较高峰值信噪比，可能是因为其生成模型直接在图像域中进行采样；而在我们的方法中，VQ-VAE 压缩可能会对指标产生不利影响。

我们将方法与“BASIS NCSN”进行比较，后者在 CelebA 上使用预训练的 NCSN 模型（Song 和 Ermon 2019）。在这种情况下，我们采用 FID 指标（Heusel 等人 2017）而非峰值信噪比进行评估，因为对于比 MNIST 更具多样性的数据集，源分离可能是一个欠定任务（Jayaram 和 Thickstun 2020）：语义上良好的分离结果可能获得较低的峰值信噪比，因为生成模型可能会改变颜色等特征和线索（VQ-GAN 解码器会放大这一影响）。FID 指标能更好地量化分离结果是否属于源的分布。我们使用带有 $k = 32$ 的 top-k 采样器，对验证集中的图像两两组合得到的 10,000 个混合信号进行测试。我们将似然项乘以 $\lambda = 3$ 进行缩放。文献中已有公认结论：在不同数据集上，基于分数的模型在 FID 指标上优于自回归模型（Dockhorn、Vahdat 和 Kreis 2021）；然而，我们的方法与自回归模型结合后，相较于基于分数的“BASIS NCSN”仍表现出竞争力。

定性结果 为了展示*LASS*在无需任何修改即可灵活使用现有模型方面的能力，我们利用了在 CelebA-HQ 和 ImageNet 上的预训练检查点。

	方法	Time
MNIST	***LASS*（我们的方法）**	4.49 秒 ± 0.27 秒
	BASIS NCSN	53.34 秒 ± 0.51 秒
SLAKH	***LASS*（我们的方法）**	1.33 分钟 ± 0.87 秒
	PnF	42.29 分钟 ± 1.08 秒

表4：计算一次分离的推理速度比较。为估计方差，我们在 MINST 上重复推理 10 次，在 SLAKH 上重复 3 次。我们将 SLAKH 上时长为 3 秒的混合信号纳入考量。

分离方法	Avg	鼓	Bass
rPCA	0.82	0.60	1.05
ICA	-1.26	-0.99	-1.53
HPSS	-0.45	-0.56	-0.33
REPET	1.04	0.53	1.54
FT2D	0.95	0.59	1.31
***LASS*（我们的方法）**	4.86	4.73	4.98
Demucs	5.39	5.42	5.36
Conv-Tasnet	5.47	5.51	5.43

表5：在 SLAKH 的“鼓”和“贝斯”类别上与其他源分离方法的比较。结果以信号失真比（dB）报告（越高越好）。表格下半部分展示了有监督方法。此处“平均值”指两个类别结果的均值。

在这种情况下，我们只学习似然张量 P。我们在图1中展示了精选结果列表，并在配套网站上提供了更全面的列表。据我们所知，我们的方法是首个能够扩展至 256×256 分辨率的方法，并且可以与更强大的潜在自回归模型配合使用，而无需重新训练（对于非常大的模型而言，重新训练十分繁琐）。因此，最终用户无需拥有用于训练这些大型模型的大量计算资源，也能进行生成式分离。

音乐源分离

我们在 SLAKH2100 数据集（Manilow 等，2019）上针对音乐源分离任务进行了实验。该数据集包含 2100 首歌曲，分离出的源属于 34 种乐器类别，混合信号总时长为 145 小时。我们重点研究“鼓”和“贝斯”两个数据类别，音轨采样率为 22 kHz。我们使用 Dhariwal 等人（2020）公开的 VQ-VAE 模型检查点，利用其在量化域中对音频数据进行建模的表达能力。鉴于该模型是在 44 kHz 下训练的，我们先对输入数据进行线性上采样，再将输出降采样回 22 kHz。对于两个自回归先验，我们训练了两个 Transformer 模型，一个用于“鼓”，另一个用于“贝斯”，并学习 VQ-VAE 上的似然函数（统计数据见表1）。我们将 *LASS* 与一组无监督盲源分离方法进行比较——“rPCA”（Huang 等，2012）、“ICA”（Hyv¨arinenand Oja，2000）、“HPSS”（Rafii 和 Pardo，2012）、“FT2D”（Seetharaman、Pishdadian 和 Pardo，2017）——并与两种有监督基线方法 Demucs（D´efossez 等，2019）和 Conv-Tasnet（Luo 和 Mesgarani，2019）进行比较，采用信号失真比（dB）作为评估指标，并使用 museval 库计算（St¨oter、Liutkus 和 Ito，2018）。为评估这些方法，我们从“鼓”和“贝斯”类别的测试划分中选取 900 段、每段 3 秒的音乐片段，并将它们组合成 450 个混合信号。验证数据集也以类似方式构建（使用不同的音乐片段）。在采样策略方面，我们采用束搜索，因为它在包含 50 个混合信号的验证集上效果最佳（表3），使用 $B = 100$ beams。评估结果见表5：*LASS* 的表现明显优于所有无监督盲源分离基线方法，且与采用有监督方法所得结果相当。此外，我们将 *LASS*与生成式源分离方法“PnF”（Ja-yaram 和 Thickstun，2021）进行速度比较，评估分离一个以 22 kHz 采样的 3 秒混合信号所需的时间（“PnF”中为钢琴与人声）。表4结果显示，*LASS* 速度显著更快，因此可应用于更实际的推理场景。

局限性

本文将分析范围限定为两个源的分离。尽管这种设置很常见，尤其是在图像分离中（Jayaram 和 Thickstun，2021年；Halperin、Ephrat 和 Hoshen，2019），但处理多个源也是未来可能的研究方向。在我们的框架下，这需要增加离散分布的维度（包括先验和似然函数）。为缓解这一问题，可以采用递归分离等技术（Takahashi 等，2019）。

所提方法的另一项局限是式（4）中采用的局部性假设。颜色化和超分辨率等不同任务需要更大的条件上下文，以及能够在全局上下文中聚合潜在编码的更新颖的量化方案（在 VQ-VAE 的编码器和解码器中使用自注意力）（Yu 等，2021）。采用针对潜在通道进行量化的 VQ-VAE（Xu 等，2021），并结合参数化似然函数，或许能够解决这一局限，同时仍保留本文所提出的 VQ-VAE、先验与似然函数之间灵活的分离方式。

结论

本文提出了 *LASS* ，这是一种针对潜在自回归模型的源分离方法，不会改变先验的结构。我们在不同数据集上测试了该方法，结果表明，其性能可与最先进的方法相媲美，同时具有更好的可扩展性，推理速度也更快。此外，我们还展示了分辨率高于竞争方法所提供结果的定性结果。我们相信，我们的方法将受益于更新一代自回归模型带来的质量提升，从而改善定量指标和感知结果。

致谢

感谢 Marco Fumero 帮助计算音频场景下的盲式无监督基线指标。本研究获得了编号为 802554（SPEC-GEO）的欧洲研究委员会（ERC）资助，以及威尼斯卡福斯卡里大学 DAIS 提供的 IRIDE 资助。

参考文献

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; 等. 2020年. 语言模型是少样本学习器。*NeurIPS* 论文集, 33: 1877–1901. Castellon, R.; Donahue, C.; and Liang, P. 2021年. 编码音频语言建模学会了对音乐信息检索有用的表示。*arXiv* 预印本 *arXiv:2107.05677*. Comon, P. 1994. 独立成分分析：一个新概念？信号处理. D´efossez, A. 2021年. 混合频谱图与波形源分离。载于*ISMIR 2021* 音乐源分离研讨会论文集. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei- Fei, L. 2009. ImageNet：大型层次化图像数据库。载于*CVPR* 论文集, 248–255. Dhariwal, P.; Jun, H.; Payne, C.; Kim, J. W.; Radford, A.; and Sutskever, I. 2020年. Jukebox：音乐生成模型。arXiv:2005.00341. Dockhorn, T.; Vahdat, A.; and Kreis, K. 2021年. 基于分数的生成建模与临界阻尼朗之万扩散。*ArXiv*, abs/2112.07068. Dovrat, S.; Nachmani, E.; and Wolf, L. 2021年. 采用最优置换训练的多说话人单通道语音分离。载于*Interspeech*. D´efossez, A.; Usunier, N.; Bottou, L.; and Bach, F. 2019. 波形域音乐源分离。*arXiv:1911.13254 [cs, eess, stat]*. ArXiv: 1911.13254. Esser, P.; Rombach, R.; and Ommer, B. 2021年. 驯服 Transformer 以实现高分辨率图像合成。载于*CVPR* 论文集, 12873–12883. Goodfellow, I.; Bengio, Y.; and Courville, A. 2016. 深度学习. MIT Press. http://www.deeplearningbook.org. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. 生成对抗网络。*NIPS* 论文集, 27. Gus´o, E.; Pons, J.; Pascual, S.; and Serr`a, J. 2022. 音乐源分离的损失函数与评估指标。载于*ICASSP* 论文集, 306–310. Halperin, T.; Ephrat, A.; and Hoshen, Y. 2019. 已观测分布与未观测分布的神经分离。第 *36* 届国际机器学习大会，*ICML 2019*, 2019-June: 4548–4557. Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. 通过交叉注意力控制进行 Prompt-to-Prompt 图像编辑。*arXiv* 预印本 *arXiv:2208.01626*。Heusel, M.；Ramsauer, H.；Unterthiner, T.；Nessler, B.；和 Hochreiter, S. 2017年。通过双时间尺度更新规则训练的生成对抗网络收敛至局部纳什均衡。载于 *NeurIPS* 论文集，第 30 卷。Ho, J.；Jain, A.；和 Abbeel, P. 2020。去噪扩散概率模型。*NeurIPS* 论文集，33：6840–6851。Holtzman, A.；Buys, J.；Du, L.；Forbes, M.；和 Choi, Y. 2020。神经文本退化的奇特现象。载于*ICLR* 论文集。Hor´e, A.；和 Ziou, D. 2010。图像质量指标：峰值信噪比与结构相似性。载于*ICPR* 论文集，2366–2369。Huang, P.-S.；Chen, S. D.；Smaragdis, P.；和 Hasegawa-Johnson, M. 2012。使用鲁棒主成分分析从单声道录音中分离歌声。载于*ICASSP* 论文集，57–60。IEEE。Huang, P.-S.；Kim, M.；Hasegawa-Johnson, M. A.；和 Smaragdis, P. 2014。使用深度循环神经网络从单声道录音中分离歌声。载于 *ISMIR* 论文集。Hyv¨arinen, A.；和 Oja, E. 2000。独立成分分析：算法与应用。神经网络，13(4-5)：411–430。Jayaram, V.；和 Thickstun, J. 2020。利用深度生成先验进行源分离。载于*ICML* 论文集，PMLR。Jayaram, V.；和 Thickstun, J. 2021。通过朗之万动力学从自回归模型中进行并行灵活采样。载于*ICML* 论文集，4807–4818。PMLR。Karras, T.；Aila, T.；Laine, S.；和 Lehtinen, J. 2018。渐进式增长生成对抗网络以提升质量、稳定性和多样性。载于*ICLR* 论文集。Kingma, D. P.；和 Welling, M. 2014。自动编码变分贝叶斯。载于*ICLR* 论文集。Kong, Q.；Xu, Y.；Wang, W.；Jackson, P. J. B.；和 Plumbley, M. D. 2019。利用生成对抗网络进行单通道信号分离与反卷积。载于*IJCAI* 论文集，2747–2753。AAAI出版社。ISBN 9780999241141。Kool, W.；van Hoof, H.；和 Welling, M. 2020。用于无放回采样的祖先式 Gumbel-Top-k 采样。机器学习研究杂志，21(47)：1–36。Lecun, Y.；Bottou, L.；Bengio, Y.；和 Haffner, P. 1998。应用于文档识别的基于梯度的学习。*IEEE* 论文集，86(11)：2278–2324。Liu, J.-Y.；和 Yang, Y.-H. 2018。用于音乐源分离的去噪自编码器、循环跳跃连接与残差回归。arXiv:1807.01898。Liu, Z.；Luo, P.；Wang, X.；和 Tang, X. 2015。在自然环境中深度学习人脸属性。载于*ICCV* 论文集。Llu´ıs, F.；Pons, J.；和 Serra, X. 2019。端到端音乐源分离：波形域中是否可行？载于*INTERSPEECH*，4619–4623。Luo, Y.；和 Mesgarani, N. 2019。Conv-TasNet：在语音分离方面超越理想时频幅度掩蔽。*IEEE/ACM* 音频、语音与语言处理汇刊，27(8)：1256–1266。

Manilow, E.；Wichern, G.；Seetharaman, P.；和 Le Roux, J. 2019。从 Slakh 中切分音乐源分离：用于研究训练数据质量与数量影响的数据集。载于*IEEE* 音频与声学信号处理应用研讨会（*WASPAA*）论文集。IEEE。Narayanaswamy, V.；Thiagarajan, J. J.；Anirudh, R.；和 Spanias, A. 2020。利用生成式先验进行无监督音频源分离。arXiv:2005.13769。Nugraha, A. A.；Liutkus, A.；和 Vincent, E. 2016。利用深度神经网络进行多通道音频源分离。*IEEE/ACM* 音频、语音与语言处理汇刊，24(9)：1652–1664。Parisi, G. 1981年。关联函数与计算机模拟。《核物理 *B*》，180(3)：378–384。Postolache, E.；Pons, J.；Pascual, S.；和 Serr`a, J. 2022。用于通用声音分离的对抗式置换不变训练。*arXiv* 预印本 *arXiv:2210.12108*。Rafii, Z.；和 Pardo, B. 2012。重复模式提取技术（REPET）：一种简单的音乐/语音分离方法。*IEEE* 音频、语音与语言处理汇刊，21(1)：73–84。Ramesh, A.；Pavlov, M.；Goh, G.；Gray, S.；Voss, C.；Radford, A.；Chen, M.；和 Sutskever, I. 2021。零样本文本到图像生成。载于*ICML* 论文集，8821–8831。PMLR。Razavi, A.；van den Oord, A.；和 Vinyals, O. 2019。利用 VQ-VAE-2 生成多样且高保真的图像。载于*NeurIPS* 论文集。Reddy, D. R.；等。1977。语音理解系统：五年研究工作的成果总结。计算机科学系。卡内基梅隆大学，宾夕法尼亚州匹兹堡，17：138。Roweis, S. T. 2000。单麦克风源分离。载于*NIPS* 论文集。Salimans, T.；Karpathy, A.；Chen, X.；和 Kingma, D. P. 2017。Pixelcnn++：通过离散化逻辑斯蒂混合似然及其他改进来提升 PixelCNN。*arXiv* 预印本 *arXiv:1701.05517*。Sanh, V.；Webson, A.；Raffel, C.；等。2022。多任务提示训练实现零样本任务泛化。载于*ICLR* 论文集。Seetharaman, P.；Pishdadian, F.；和 Pardo, B. 2017。使用二维傅里叶变换进行音乐/语音分离。载于*2017* 年 *IEEE* 音频与声学信号处理应用研讨会（*WASPAA*），36–40。IEEE。Smaragdis, P.；F´evotte, C.；Mysore, G. J.；Mohammadiha, N.；和 Hoffman, M. 2014。使用非负分解进行静态与动态源分离：统一视角。《*IEEE* 信号处理杂志》，31(3)：66–75。Song, Y.；和 Ermon, S. 2019。通过估计数据分布的梯度进行生成建模。载于《神经信息处理系统进展》，11895–11907。Song, Y.；Sohl-Dickstein, J.；Kingma, D. P.；Kumar, A.；Ermon, S.；和 Poole, B. 2021。通过随机微分方程进行基于得分的生成建模。载于*ICLR* 论文集

St¨oter, F.-R.; Liutkus, A.; 和 Ito, N. 2018. 2018年信号分离评测竞赛。载于*LVA/ICA* 会议论文集，293–305。Subakan, Y. C.; 和 Smaragdis, P. 2018. 生成对抗式源分离。载于*ICASSP* 会议论文集，26–30。IEEE。Takahashi, N.; Goswami, N.; 和 Mitsufuji, Y. 2018. Mm-denselstm：一种用于音频源分离的卷积神经网络与循环神经网络的高效组合。载于*IWAENC* 会议论文集，106–110。Takahashi, N.; Parthasaarathy, S.; Goswami, N.; 和 Mitsufuji, Y. 2019. 面向说话人数未知场景的递归语音分离。*arXiv* 预印本 *arXiv:1904.03065*。Uhlich, S.; Giron, F.; 和 Mitsufuji, Y. 2015. 基于深度神经网络的音乐乐器提取。载于*ICASSP* 会议论文集。van den Oord, A.; Kalchbrenner, N.; Espeholt, L.; Vinyals, O.; Graves, A.; 等。2016. 使用 PixelCNN 解码器进行条件图像生成。*NeurIPS* 会议论文集，29。van de n Oord, A.; Vinyals, O.; 和 Kavukcuoglu, K. 2017. 神经离散表征学习。载于*NeurIPS* 会议论文集。Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; 和 Polosukhin, I. 2017. 注意力就是你所需的一切。*NeurIPS* 会议论文集，30。Wei, J.; Bosma, M.; Zhao, V. Y.; Guu, K.; Yu, A. W.; Lester, B.; Du, N.; Dai, A. M.; 和 Le, Q. V. 2021. 微调语言模型是零样本学习者。*CoRR*，abs/2109.01652。Wisdom, S.; Erdogan, H.; Ellis, D. P. W.; Serizel, R.; Turpault, N.; Fonseca, E.; Sala mon, J.; Seetharaman, P.; 和 Hershey, J. R. 2021. 关于免费通用声音分离数据的种种讨论究竟为何？载于*ICASSP* 会议论文集，186–190。Wisdom, S.; Tzinis, E.; Erdogan, H.; Weiss, R. J.; Wilson, K. W.; 和 Hershey, J. R. 2020. 使用混合不变训练进行无监督声音分离。Xu, Y.; Song, Y.; Garg, S.; Gong, L.; Shu, R.; Grover, A.; 和 Ermon, S. 2021. 通过有序自编码对自回归模型进行任意时刻采样。*arXiv* 预印本 *arXiv:2102.11495*。Yang, H.; Lin, J.; Yang, A.; Wang, P.; Zhou, C.; 和 Yang, H. 2022. 面向生成式多模态预训练模型的提示调优。Yu, J.; Li, X.; Koh, J. Y.; Zhang, H.; Pang, R.; Qin, J.; Ku, A.; Xu, Y.; Baldridge, J.; 和 Wu, Y. 2021. 改进 VQGAN 的向量量化图像建模。*arXiv* 预印本 *arXiv:2110.04627*。Yu, J.; Xu, Y.; Koh, J. Y.; Luong, T.; Baid, G.; Wang, Z.; Vasudevan, V.; Ku, A.; Yang, Y.; Ayan, B. K.; 等。2022. 扩展自回归模型以生成内容丰富的文本图像。*arXiv* 预印本 *arXiv:2206.10789*。

.