无分类器引导的扩散模型

Jonathan Ho & Tim Salimans
Google Research, Brain team
{jonathanho,salimans}@google.com

摘要

分类器引导（classifier guidance）是一种新近提出的方法，用于在训练后权衡条件扩散模型的模式覆盖与样本保真度，其精神与在其他类型生成模型中使用的低温采样或截断类似。分类器引导将扩散模型的分数估计与图像分类器的梯度相结合，因此需要训练一个独立于扩散模型的图像分类器。这也引出了一个问题：是否可以在没有分类器的情况下进行引导？我们证明，纯生成模型确实可以在没有此类分类器的情况下进行引导：在我们称之为无分类器引导（classifier-free guidance）的方法中，我们联合训练一个条件扩散模型和一个无条件扩散模型，并将得到的条件与无条件分数估计相结合，以获得与使用分类器引导类似的样本质量与多样性之间的权衡。

1 引言

扩散模型最近作为一种富有表达力和灵活性的生成模型家族崭露头角，在图像和音频合成任务上提供了具有竞争力的样本质量和似然分数 (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Ho et al., 2020; Song et al., 2021b; Kingma et al., 2021; Song et al., 2021a)。这些模型在音频合成方面取得了与自回归模型相媲美的质量，同时推理步骤大幅减少 (Chen et al., 2021; Kong et al., 2021)，并且在 ImageNet 生成结果上，就 FID 分数和分类准确率分数而言，超越了 BigGAN-deep (Brock et al., 2019) 和 VQ-VAE-2 (Razavi et al., 2019) (Ho et al., 2021; Dhariwal & Nichol, 2021)。

图 1: 无分类器引导在 64x64 ImageNet 扩散模型中对 malamute 类别的效果。从左到右：无分类器引导程度逐渐增加，最左侧为无引导采样。

本文为 AI 辅助翻译版本，仅供学习交流；引用请以英文原文为准：arXiv:2207.12598。
本文的短版本曾发表于 NeurIPS 2021 深度生成模型及其下游应用研讨会：https://openreview.net/pdf?id=qw8AKxfYbI

图 2: 引导对三个高斯混合分布的影响，每个混合分量代表以类别为条件的数据。最左侧的图是无引导的边缘密度。从左到右是归一化引导条件分布的混合密度，引导强度逐渐增加。

Dhariwal & Nichol (2021) 提出了*分类器引导*，一种利用额外训练好的分类器来提升扩散模型样本质量的技术。在分类器引导出现之前，人们不知道如何从扩散模型中生成类似截断 BigGAN (Brock et al., 2019) 或低温 Glow (Kingma & Dhariwal, 2018) 所生成的" 低温" 样本：朴素尝试，如缩放模型分数向量或降低扩散采样过程中添加的高斯噪声量，都是无效的 (Dhariwal & Nichol, 2021)。分类器引导则是将扩散模型的分数估计与分类器输入的对数概率梯度混合。通过改变分类器梯度的强度，Dhariwal & Nichol可以以类似于改变 BigGAN 截断参数的方式，在 Inception 分数 (Salimans et al., 2016) 和 FID 分数 (Heusel et al., 2017)（或精确率和召回率）之间进行权衡。

我们感兴趣的是，分类器引导是否可以在没有分类器的情况下进行。分类器引导使扩散模型的训练流程变得复杂，因为它需要训练一个额外的分类器，并且这个分类器必须在带噪声的数据上训练，因此通常无法直接插入预训练的分类器。此外，由于分类器引导在采样时将分数估计与分类器梯度混合，分类器引导的扩散采样可以被解释为试图用基于梯度的对抗攻击来迷惑图像分类器。这引出了一个问题：分类器引导之所以能成功提升基于分类器的指标（如 FID 和 Inception 分数），是否仅仅因为它对这类分类器具有对抗性。沿着分类器梯度的方向迈步也与 GAN 训练有一些相似之处，特别是与非参数生成器；这也引出了一个问题：分类器引导的扩散模型在基于分类器的指标上表现良好，是否因为它们开始类似于 GAN，而 GAN 已知在这类指标上表现良好。

为解决这些问题，我们提出了*无分类器引导*，这是一种完全避免使用任何分类器的引导方法。与其沿着图像分类器梯度的方向采样，无分类器引导而是混合条件扩散模型和联合训练的无条件扩散模型的分数估计。通过扫描混合权重，我们获得了与分类器引导类似的 FID/IS 权衡。我们的无分类器引导结果表明，纯生成扩散模型能够合成与其他类型生成模型相媲美的极高保真度样本。

2 背景

我们在连续时间中训练扩散模型 (Song et al., 2021b; Chen et al., 2021; Kingma et al., 2021)：令 $\mathbf{x} \sim p(\mathbf{x})$ 和 $\mathbf{z} = \{\mathbf{z}_\lambda \mid \lambda \in [\lambda_{\min}, \lambda_{\max}]\}$，其中超参数 $\lambda_{\min} < \lambda_{\max} \in \mathbb{R}$，前向过程 $q(\mathbf{z}|\mathbf{x})$ 是保持方差的马尔可夫过程 (Sohl-Dickstein et al., 2015)：

$$q(\mathbf{z}_\lambda|\mathbf{x}) = \mathcal{N}(\alpha_\lambda \mathbf{x}, \sigma_\lambda^2 \mathbf{I}), \text{ where } \alpha_\lambda^2 = 1/(1+e^{-\lambda}),\ \sigma_\lambda^2 = 1-\alpha_\lambda^2 \tag{1}$$

$$q(\mathbf{z}_\lambda|\mathbf{z}_{\lambda'}) = \mathcal{N}((\alpha_\lambda/\alpha_{\lambda'})\mathbf{z}_{\lambda'}, \sigma_{\lambda|\lambda'}^2 \mathbf{I}), \text{ where } \lambda < \lambda',\ \sigma_{\lambda|\lambda'}^2 = (1-e^{\lambda-\lambda'})\sigma_\lambda^2 \tag{2}$$

我们将使用记号 $p(\mathbf{z})$（或 $p(\mathbf{z}_\lambda)$）来表示当 $\mathbf{x} \sim p(\mathbf{x})$ 且 $\mathbf{z} \sim q(\mathbf{z}|\mathbf{x})$ 时 $\mathbf{z}$（或 $\mathbf{z}_\lambda$）的边缘分布。注意 $\lambda = \log \alpha_\lambda^2/\sigma_\lambda^2$，因此 λ 可以解释为 $\mathbf{z}_\lambda$ 的对数信噪比，前向过程沿 λ 减小的方向运行。

给定 $\mathbf{x}$，前向过程可以通过反向转移来描述 $q(\mathbf{z}_{\lambda'}|\mathbf{z}_\lambda, \mathbf{x}) = \mathcal{N}(\tilde{\boldsymbol{\mu}}_{\lambda'|\lambda}(\mathbf{z}_\lambda, \mathbf{x}), \tilde{\sigma}^2_{\lambda'|\lambda}\mathbf{I})$，其中

$$\tilde{\boldsymbol{\mu}}_{\lambda'|\lambda}(\mathbf{z}_\lambda, \mathbf{x}) = e^{\lambda-\lambda'}(\alpha_{\lambda'}/\alpha_\lambda)\mathbf{z}_\lambda + (1-e^{\lambda-\lambda'})\alpha_{\lambda'}\mathbf{x}, \quad \tilde{\sigma}^2_{\lambda'|\lambda} = (1-e^{\lambda-\lambda'})\sigma^2_{\lambda'} \tag{3}$$

反向过程生成模型从 $p_\theta(\mathbf{z}_{\lambda_{\min}}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ 开始。我们指定转移：

$$p_\theta(\mathbf{z}_{\lambda'}|\mathbf{z}_\lambda) = \mathcal{N}(\tilde{\boldsymbol{\mu}}_{\lambda'|\lambda}(\mathbf{z}_\lambda, \mathbf{x}_\theta(\mathbf{z}_\lambda)), (\tilde{\sigma}^2_{\lambda'|\lambda})^{1-v}(\sigma^2_{\lambda|\lambda'})^v) \tag{4}$$

在采样过程中，我们沿递增序列 $\lambda_{\min} = \lambda_1 < \cdots < \lambda_T = \lambda_{\max}$ 应用此转移，共 T 个时间步；换句话说，我们遵循 Sohl-Dickstein et al. (2015); Ho et al. (2020) 的离散时间祖先采样器。如果模型 $\mathbf{x}_\theta$ 是正确的，那么当 $T \to \infty$ 时，我们获得一个 SDE 的样本，其样本路径服从 $p(\mathbf{z})$ (Song et al., 2021b)，我们用 $p_\theta(\mathbf{z})$ 表示连续时间模型分布。方差是 $\tilde{\sigma}^2_{\lambda'|\lambda}$ 和 $\sigma^2_{\lambda|\lambda'}$ 的对数空间插值，如 Nichol & Dhariwal (2021) 所建议；我们发现使用常数超参数 v 比学习的 $\mathbf{z}_\lambda$ 相关的 v 更有效。注意当 $\lambda' \to \lambda$ 时，方差简化为 $\tilde{\sigma}^2_{\lambda'|\lambda}$，因此 v 仅在采样时使用非无穷小时间步时才有效，正如实践中那样。

反向过程均值来自一个估计 $\mathbf{x}_\theta(\mathbf{z}_\lambda) \approx \mathbf{x}$，将其代入 $q(\mathbf{z}_{\lambda'}|\mathbf{z}_\lambda, \mathbf{x})$ (Ho et al., 2020; Kingma et al., 2021)（$\mathbf{x}_\theta$ 也接收 λ 作为输入，但我们为保持记号简洁而省略）。我们用 $\boldsymbol{\epsilon}$-预测 (Ho et al., 2020) 来参数化 $\mathbf{x}_\theta$：$\mathbf{x}_\theta(\mathbf{z}_\lambda) = (\mathbf{z}_\lambda - \sigma_\lambda\boldsymbol{\epsilon}_\theta(\mathbf{z}_\lambda))/\alpha_\lambda$，并在目标上训练

$$\mathbb{E}_{\boldsymbol{\epsilon},\lambda}\left[\|\boldsymbol{\epsilon}_\theta(\mathbf{z}_\lambda) - \boldsymbol{\epsilon}\|_2^2\right] \tag{5}$$

其中 $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$，$\mathbf{z}_\lambda = \alpha_\lambda\mathbf{x} + \sigma_\lambda\boldsymbol{\epsilon}$，且 λ 从 $[\lambda_{\min}, \lambda_{\max}]$ 上的分布 $p(\lambda)$ 中抽取。该目标是多噪声尺度下的去噪分数匹配 (Vincent, 2011; Hyvärinen & Dayan, 2005) (Song & Ermon, 2019)，当 $p(\lambda)$ 为均匀分布时，该目标与隐变量模型 $\int p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z})d\mathbf{z}$ 的边缘对数似然的变分下界成比例，忽略未指定的解码器 $p_\theta(\mathbf{x}|\mathbf{z})$ 和 $\mathbf{z}_{\lambda_{\min}}$ 处先验的项 (Kingma et al., 2021)。

如果 $p(\lambda)$ 不是均匀分布，该目标可以解释为加权变分下界，其权重可以调整以优化样本质量 (Ho et al., 2020; Kingma et al., 2021)。我们使用受 Nichol & Dhariwal (2021) 离散时间余弦噪声调度启发的 $p(\lambda)$：我们通过 $\lambda = -2\log\tan(au+b)$ 采样 λ，其中 $u \in [0,1]$ 均匀分布，$b = \arctan(e^{-\lambda_{\max}/2})$，$a = \arctan(e^{-\lambda_{\min}/2}) - b$。这表示一个双曲正割分布，修改后支持在有界区间上。对于有限时间步生成，我们使用对应于均匀间隔的 $u \in [0,1]$ 的 λ 值，最终生成的样本为 $\mathbf{x}_\theta(\mathbf{z}_{\lambda_{\max}})$。

由于对所有 λ 而言，$\boldsymbol{\epsilon}_\theta(\mathbf{z}_\lambda)$ 的损失是去噪分数匹配，我们的模型学到的分数 $\boldsymbol{\epsilon}_\theta(\mathbf{z}_\lambda)$ 估计了噪声数据 $\mathbf{z}_\lambda$ 分布的对数密度梯度，即 $\boldsymbol{\epsilon}_\theta(\mathbf{z}_\lambda) \approx -\sigma_\lambda\nabla_{\mathbf{z}_\lambda}\log p(\mathbf{z}_\lambda)$；但请注意，由于我们使用无约束神经网络来定义 $\boldsymbol{\epsilon}_\theta$，因此不一定存在任何标量势，其梯度为 $\boldsymbol{\epsilon}_\theta$。从学习到的扩散模型中采样类似于使用 Langevin 扩散从一系列分布 $p(\mathbf{z}_\lambda)$ 中采样，这些分布收敛到原始数据 $\mathbf{x}$ 的条件分布 $p(\mathbf{x})$。

在条件生成建模的情况下，数据 $\mathbf{x}$ 与条件信息 $\mathbf{c}$ 联合抽取，即用于类别条件图像生成的类别标签。对模型的唯一修改是反向过程函数逼近器接收 $\mathbf{c}$ 作为输入，如 $\boldsymbol{\epsilon}_\theta(\mathbf{z}_\lambda, \mathbf{c})$。

3 引导

某些生成模型（如 GAN 和基于流的模型）的一个有趣特性是能够在采样时通过降低噪声输入的方差或范围来进行截断或低温采样。预期效果是降低样本多样性，同时提高每个单独样本的质量。例如，BigGAN (Brock et al., 2019) 中的截断在低截断量和高截断量下分别产生 FID 分数和 Inception 分数之间的权衡曲线。Glow (Kingma & Dhariwal, 2018) 中的低温采样具有类似效果。

不幸的是，在扩散模型中直接实现截断或低温采样的尝试是无效的。例如，缩放模型分数或降低反向过程中高斯噪声的方差会导致扩散模型生成模糊、低质量的样本 (Dhariwal & Nichol, 2021)。

3.1 分类器引导

为了在扩散模型中获得类似截断的效果，Dhariwal & Nichol (2021) 引入了分类器引导，其中扩散分数 $\boldsymbol{\epsilon}_\theta(\mathbf{z}_\lambda, \mathbf{c}) \approx -\sigma_\lambda \nabla_{\mathbf{z}_\lambda} \log p(\mathbf{z}_\lambda|\mathbf{c})$ 被修改为包含辅助分类器模型 $p_\theta(\mathbf{c}|\mathbf{z}_\lambda)$ 的对数似然梯度，如下所示：

$$\tilde{\boldsymbol{\epsilon}}_\theta(\mathbf{z}_\lambda, \mathbf{c}) = \boldsymbol{\epsilon}_\theta(\mathbf{z}_\lambda, \mathbf{c}) - w\sigma_\lambda \nabla_{\mathbf{z}_\lambda} \log p_\theta(\mathbf{c}|\mathbf{z}_\lambda) \approx -\sigma_\lambda \nabla_{\mathbf{z}_\lambda} [\log p(\mathbf{z}_\lambda|\mathbf{c}) + w \log p_\theta(\mathbf{c}|\mathbf{z}_\lambda)],$$

其中 w 是控制分类器引导强度的参数。这个修改后的分数 $\tilde{\boldsymbol{\epsilon}}_\theta(\mathbf{z}_\lambda, \mathbf{c})$ 在从扩散模型采样时替代 $\boldsymbol{\epsilon}_\theta(\mathbf{z}_\lambda, \mathbf{c})$，从而近似地从以下分布中采样

$$\tilde{p}_\theta(\mathbf{z}_\lambda|\mathbf{c}) \propto p_\theta(\mathbf{z}_\lambda|\mathbf{c}) p_\theta(\mathbf{c}|\mathbf{z}_\lambda)^w.$$

其效果是提升那些分类器 $p_\theta(\mathbf{c}|\mathbf{z}_\lambda)$ 赋予正确标签高似然的数据的概率：能够被很好分类的数据在感知质量的 Inception 分数 (Salimans et al., 2016) 上得分很高，而该指标天生就奖励生成模型做到这一点。Dhariwal & Nichol 因此发现，通过设置 $w > 0$ 可以提高扩散模型的 Inception 分数，但代价是样本多样性降低。

Figure 2 展示了数值求解的引导 $\tilde{p}_\theta(\mathbf{z}_\lambda|\mathbf{c}) \propto p_\theta(\mathbf{z}_\lambda|\mathbf{c}) p_\theta(\mathbf{c}|\mathbf{z}_\lambda)^w$ 在一个三类别二维玩具示例上的效果，其中每个类别的条件分布为各向同性高斯分布。施加引导后，各条件分布的形态明显偏离高斯。随着引导强度增大，各条件分布将概率质量推向远离其他类别、移向逻辑回归给出的高置信方向，大部分质量集中到更小的区域。这一行为可视为 ImageNet 模型上增大分类器引导强度时 Inception 分数上升、样本多样性下降的一种简化显现。

从理论上讲，对无条件模型施加权重为 $w + 1$ 的分类器引导，与对条件模型施加权重为 w 的分类器引导会得到相同结果，因为 $p_\theta(\mathbf{z}_\lambda|\mathbf{c}) p_\theta(\mathbf{c}|\mathbf{z}_\lambda)^w \propto p_\theta(\mathbf{z}_\lambda) p_\theta(\mathbf{c}|\mathbf{z}_\lambda)^{w+1}$；用分数表示即

$$\begin{aligned}\boldsymbol{\epsilon}_\theta(\mathbf{z}_\lambda) - (w+1)\sigma_\lambda \nabla_{\mathbf{z}_\lambda} \log p_\theta(\mathbf{c}|\mathbf{z}_\lambda) &\approx -\sigma_\lambda \nabla_{\mathbf{z}_\lambda} [\log p(\mathbf{z}_\lambda) + (w+1) \log p_\theta(\mathbf{c}|\mathbf{z}_\lambda)] \\ &= -\sigma_\lambda \nabla_{\mathbf{z}_\lambda} [\log p(\mathbf{z}_\lambda|\mathbf{c}) + w \log p_\theta(\mathbf{c}|\mathbf{z}_\lambda)],\end{aligned}$$

但有趣的是，Dhariwal & Nichol 在对已经具备类别条件的模型施加分类器引导时获得了最佳结果，而非对无条件模型施加引导。基于此，我们将沿用对已条件模型进行引导的设定。

3.2 无分类器引导

分类器引导虽能如截断或低温采样所预期的那样权衡 IS 与 FID，但仍依赖图像分类器的梯度；出于 Section 1 所述原因，我们希望摆脱分类器。这里我们介绍无分类器引导，它无需此类梯度即可达到相同效果。无分类器引导是修改 $\boldsymbol{\epsilon}_\theta(\mathbf{z}_\lambda, \mathbf{c})$ 的一种替代方法，作用与分类器引导相同，但不使用分类器。Algorithms 1 and 2 详细描述了无分类器引导的训练与采样过程。

我们不训练单独的分类器模型，而是选择联合训练一个无条件去噪扩散模型 $p_\theta(\mathbf{z})$ 与一个条件模型 $p_\theta(\mathbf{z}|\mathbf{c})$，前者通过得分估计器 $\boldsymbol{\epsilon}_\theta(\mathbf{z}_\lambda)$ 参数化，后者通过 $\boldsymbol{\epsilon}_\theta(\mathbf{z}_\lambda, \mathbf{c})$ 参数化。我们用同一个神经网络对两个模型进行参数化：对于无条件模型，预测得分时只需将类别标识 $\mathbf{c}$ 替换为空记号 $\varnothing$，即 $\boldsymbol{\epsilon}_\theta(\mathbf{z}_\lambda) = \boldsymbol{\epsilon}_\theta(\mathbf{z}_\lambda, \mathbf{c} = \varnothing)$。联合训练无条件模型与条件模型的方式极为简单：以某一概率 p_{uncond} 随机将 $\mathbf{c}$ 置为无条件类别标识 $\varnothing$，该概率作为超参数。（当然也

算法 1 用无分类器引导联合训练扩散模型

Require: p_{uncond}：无条件训练概率

repeat
  $(\mathbf{x}, \mathbf{c}) \sim p(\mathbf{x}, \mathbf{c})$ ▷ 从数据集采样带条件的数据
  $\mathbf{c} \leftarrow \varnothing$ with probability p_{uncond} ▷ 随机丢弃条件以进行无条件训练
  $\lambda \sim p(\lambda)$ ▷ 采样 log SNR 值
  $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  $\mathbf{z}_\lambda = \alpha_\lambda \mathbf{x} + \sigma_\lambda \boldsymbol{\epsilon}$ ▷ 将数据加噪至所采样的 log SNR 值
  Take gradient step on $\nabla_\theta \left\| \boldsymbol{\epsilon}_\theta(\mathbf{z}_\lambda, \mathbf{c}) - \boldsymbol{\epsilon} \right\|^2$ ▷ 去噪模型优化
until converged

算法 2 用无分类器引导进行条件采样

Require: w：引导强度

Require: $\mathbf{c}$：条件采样所用的条件信息

Require: $\lambda_1, \ldots, \lambda_T$：递增的 log SNR 序列，其中 $\lambda_1 = \lambda_{\min}$，$\lambda_T = \lambda_{\max}$

$\mathbf{z}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = 1, \ldots, T$ **do**
  ▷ 在 log SNR λ_t 处构造无分类器引导分数
  $\tilde{\boldsymbol{\epsilon}}_t = (1 + w)\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, \mathbf{c}) - w\boldsymbol{\epsilon}_\theta(\mathbf{z}_t)$
  ▷ 采样步（可替换为其他采样器，如 DDIM）
  $\tilde{\mathbf{x}}_t = (\mathbf{z}_t - \sigma_{\lambda_t} \tilde{\boldsymbol{\epsilon}}_t)/\alpha_{\lambda_t}$
  $\mathbf{z}_{t+1} \sim \mathcal{N}(\tilde{\boldsymbol{\mu}}_{\lambda_{t+1}|\lambda_t}(\mathbf{z}_t, \tilde{\mathbf{x}}_t), (\tilde{\sigma}^2_{\lambda_{t+1}|\lambda_t})^{1-v} (\sigma^2_{\lambda_t|\lambda_{t+1}})^v)$ if $t < T$ else $\mathbf{z}_{t+1} = \tilde{\mathbf{x}}_t$
end for
return $\mathbf{z}_{T+1}$

可以分别训练两个模型而非联合训练，但我们选择联合训练，因为它实现起来极其简单，既不使训练流程复杂化，也不增加总参数量。）随后，我们使用条件得分估计与无条件得分估计的如下线性组合进行采样：

$$\tilde{\boldsymbol{\epsilon}}_\theta(\mathbf{z}_\lambda, \mathbf{c}) = (1 + w)\boldsymbol{\epsilon}_\theta(\mathbf{z}_\lambda, \mathbf{c}) - w\boldsymbol{\epsilon}_\theta(\mathbf{z}_\lambda) \tag{6}$$

Eq. (6) 中不存在分类器梯度，因此沿着 $\tilde{\boldsymbol{\epsilon}}_\theta$ 方向迈进一步不能被解释为对图像分类器进行基于梯度的对抗攻击。此外，$\tilde{\boldsymbol{\epsilon}}_\theta$ 由得分估计构成，由于使用了无约束神经网络，这些得分估计构成非保守向量场，因此一般而言不存在标量势（如分类器对数似然），使得 $\tilde{\boldsymbol{\epsilon}}_\theta$ 成为分类器引导得分。

尽管一般而言可能不存在一个分类器使得 Eq. (6) 是其分类器引导得分，但它实际上受到了隐式分类器 $p^i(\mathbf{c}|\mathbf{z}_\lambda) \propto p(\mathbf{z}_\lambda|\mathbf{c})/p(\mathbf{z}_\lambda)$ 梯度的启发。如果我们能获取精确的得分 $\boldsymbol{\epsilon}^*(\mathbf{z}_\lambda, \mathbf{c})$ 和 $\boldsymbol{\epsilon}^*(\mathbf{z}_\lambda)$（分别对应于 $p(\mathbf{z}_\lambda|\mathbf{c})$ 和 $p(\mathbf{z}_\lambda)$），那么这个隐式分类器的梯度将是 $\nabla_{\mathbf{z}_\lambda} \log p^i(\mathbf{c}|\mathbf{z}_\lambda) = -\frac{1}{\sigma_\lambda}[\boldsymbol{\epsilon}^*(\mathbf{z}_\lambda, \mathbf{c}) - \boldsymbol{\epsilon}^*(\mathbf{z}_\lambda)]$，而使用这个隐式分类器的分类器引导会将得分估计修改为 $\tilde{\boldsymbol{\epsilon}}^*(\mathbf{z}_\lambda, \mathbf{c}) = (1 + w)\boldsymbol{\epsilon}^*(\mathbf{z}_\lambda, \mathbf{c}) - w\boldsymbol{\epsilon}^*(\mathbf{z}_\lambda)$。注意它与 Eq. (6) 的相似性，但也要注意 $\tilde{\boldsymbol{\epsilon}}^*(\mathbf{z}_\lambda, \mathbf{c})$ 与 $\tilde{\boldsymbol{\epsilon}}_\theta(\mathbf{z}_\lambda, \mathbf{c})$ 存在根本区别。前者由缩放后的分类器梯度 $\boldsymbol{\epsilon}^*(\mathbf{z}_\lambda, \mathbf{c}) - \boldsymbol{\epsilon}^*(\mathbf{z}_\lambda)$ 构成；后者由估计值 $\boldsymbol{\epsilon}_\theta(\mathbf{z}_\lambda, \mathbf{c}) - \boldsymbol{\epsilon}_\theta(\mathbf{z}_\lambda)$ 构成，而这个表达式一般而言并不是任何分类器的（缩放）梯度，同样因为得分估计是无约束神经网络的输出。

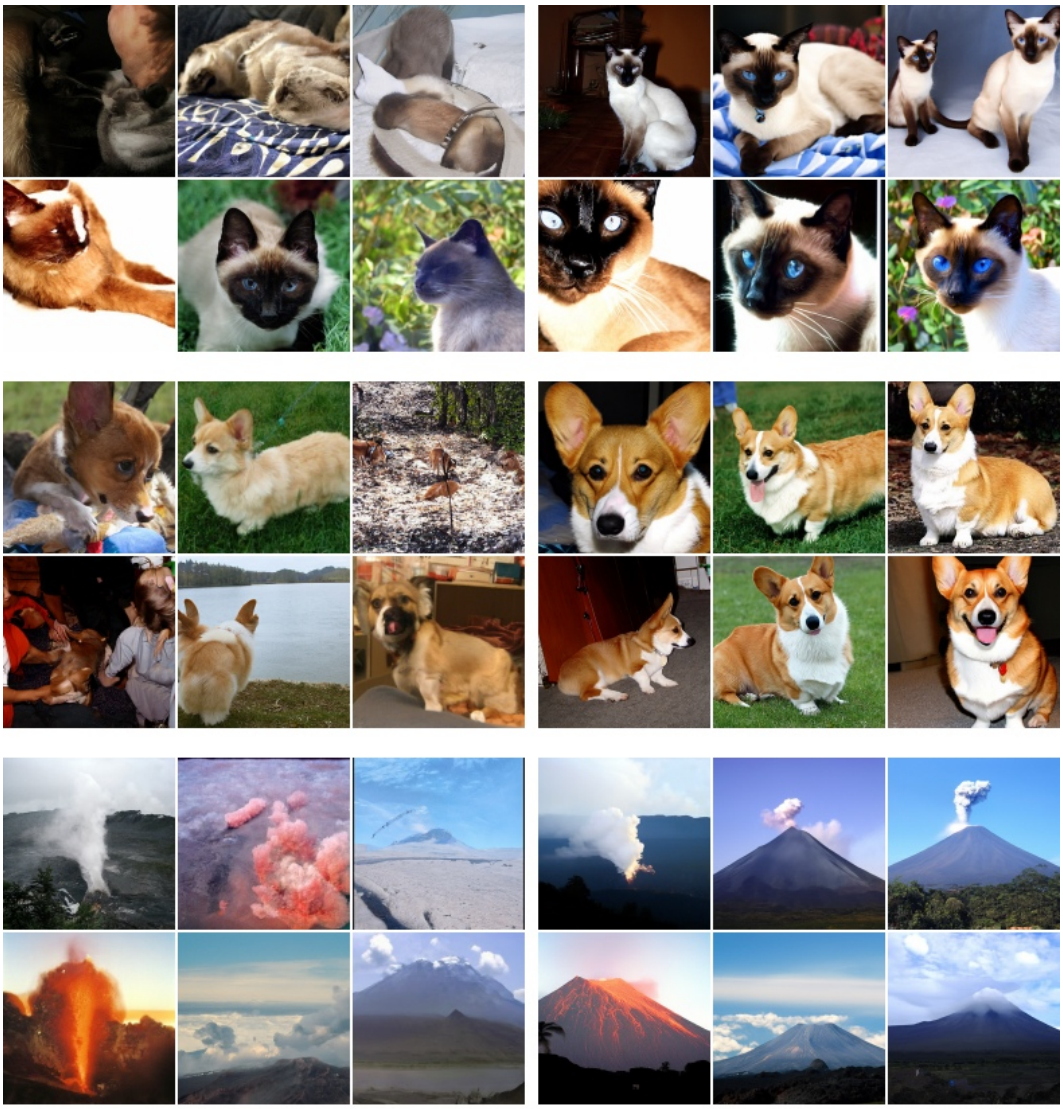


图 3: 128x128 ImageNet 上的无分类器引导。左：无引导样本；右：无分类器引导样本，$w = 3.0$。有趣的是，这类强引导样本呈现饱和色彩。更多样本见 Fig. 8。

先验地看，用贝叶斯规则反演生成模型并不一定能得到一个好的分类器来提供有用的引导信号。例如，Grandvalet & Bengio (2004) 发现判别模型通常优于从生成模型推导出的隐式分类器，即使在那些生成模型的设定与数据分布完全匹配的人工案例中。在我们这种情形下，由于模型设定可能存在偏差，通过贝叶斯规则得到的分类器可能是不一致的 (Grünwald & Langford, 2007)，我们对其性能没有任何保证。尽管如此，在 Section 4 中，我们通过实验证明无分类器引导能够以与分类器引导相同的方式权衡 FID 和 IS。在 Section 5 中，我们讨论了无分类器引导相对于分类器引导的含义。

4 实验

我们在经区域下采样的类别条件 ImageNet (Russakovsky et al., 2015) 上训练带无分类器引导的扩散模型；自 BigGAN 论文 (Brock et al., 2019) 起，这是研究 FID 与 Inception 分数之间权衡的标准设置。

实验旨在作为概念验证，说明无分类器引导能获得与分类器引导类似的 FID/IS 权衡，并理解无分类器引导的行为，而非一定要在这些基准上把样本质量指标推至最优。为此，我们采用与 Dhariwal & Nichol (2021) 的引导扩散模型相同的模型架构和超参数（除 Section 2 所述的连续时间训练外）；这些超参数是为分类器引导调优的，对无分类器引导可能并非最优。此外，由于我们将条件模型与无条件模型合并到同一架构中且不引入额外分类器，模型容量实际上低于此前的工作。尽管如此，我们的无分类器引导模型仍产生了有竞争力的样本质量指标，在某些情况下还超过了此前的工作，详见后续各小节。

4.1 改变无分类器引导强度

本小节实验性地验证本文的主要论断：无分类器引导能像分类器引导或 GAN 截断那样在 IS 与 FID 之间进行权衡。我们将所提无分类器引导应用于 64×64 和 128×128 类别条件 ImageNet 生成。Table 1 和 Fig. 4 展示了在 64×64 ImageNet 模型上扫描引导强度 w 时样本质量的变化；Table 2 和 Fig. 5 展示 128×128 模型的对应结果。我们考虑 $w \in \{0, 0.1, 0.2, \ldots, 4\}$，并对每个 w 值按 Heusel et al. (2017) 和 Salimans et al. (2016) 的流程用 50000 个样本计算 FID 和 Inception 分数。所有模型使用对数信噪比端点 $\lambda_{\min} = -20$ 和 $\lambda_{\max} = 20$。64×64 模型的采样器噪声插值系数 $v = 0.3$，训练 40 万步；128×128 模型 $v = 0.2$，训练 270 万步。

少量引导（$w = 0.1$ 或 $w = 0.3$，取决于数据集）即可获得最佳 FID 结果，而强引导（$w \geq 4$）获得最佳 IS 结果。在这两个极端之间，可以清晰看到两种感知质量指标之间的权衡：FID 随 w 单调下降，IS 随 w 单调上升。我们的结果优于 Dhariwal & Nichol (2021) 和 Ho et al. (2021)；事实上，我们的 128×128 结果是该领域文献中的最优水平。当 $w = 0.3$ 时，我们的模型在 128×128 ImageNet 上的 FID 分数超过分类器引导的 ADM-G；当 $w = 4.0$ 时，在 BigGAN-deep 处于其最佳 IS 截断水平下评估，我们的模型在 FID 和 IS 上均超过 BigGAN-deep。

Figs. 1, 3 and 6 to 8 展示了不同引导强度下模型的随机生成样本：可以清楚看到，增大无分类器引导强度会如预期那样降低样本多样性、提高单个样本的保真度。

4.2 改变无条件训练概率

无分类器引导在训练时的主要超参数是 p_{uncond}，即在联合训练条件扩散模型与无条件扩散模型时以无条件生成方式训练的概率。本小节在 64×64 ImageNet 上研究改变 p_{uncond} 对模型的影响。

Table 1 和 Fig. 4 展示了 p_{uncond} 对样本质量的影响。我们训练了 $p_{\text{uncond}} \in \{0.1, 0.2, 0.5\}$ 的模型，均训练 40 万步，并在多种引导强度下评估样本质量。我们发现，在整个 IS/FID 前沿上，$p_{\text{uncond}} = 0.5$ 始终劣于 $p_{\text{uncond}} \in \{0.1, 0.2\}$；$p_{\text{uncond}} \in \{0.1, 0.2\}$ 之间表现大致相当。

基于这些发现，我们得出结论：只需将扩散模型相对较小的一部分模型容量分配给无条件生成任务，即可产生对样本质量有效的无分类器引导得分。有趣的是，对于分类器引导，

Model	FID (↓)	IS (↑)
ADM (Dhariwal & Nichol, 2021)	2.07	-
CDM (Ho et al., 2021)	**1.48**	67.95
Ours	$p_{\text{uncond}} = 0.1/0.2/0.5$	
$w = 0.0$	1.8 / 1.8 / 2.21	53.71 / 52.9 / 47.61
$w = 0.1$	1.55 / 1.62 / 1.91	66.11 / 64.58 / 56.1
$w = 0.2$	2.04 / 2.1 / 2.08	78.91 / 76.99 / 65.6
$w = 0.3$	3.03 / 2.93 / 2.65	92.8 / 88.64 / 74.92
$w = 0.4$	4.3 / 4 / 3.44	106.2 / 101.11 / 84.27
$w = 0.5$	5.74 / 5.19 / 4.34	119.3 / 112.15 / 92.95
$w = 0.6$	7.19 / 6.48 / 5.27	131.1 / 122.13 / 102
$w = 0.7$	8.62 / 7.73 / 6.23	141.8 / 131.6 / 109.8
$w = 0.8$	10.08 / 8.9 / 7.25	151.6 / 140.82 / 116.9
$w = 0.9$	11.41 / 10.09 / 8.21	161 / 150.26 / 124.6
$w = 1.0$	12.6 / 11.21 / 9.13	170.1 / 158.29 / 131.1
$w = 2.0$	21.03 / 18.79 / 16.16	225.5 / 212.98 / 183
$w = 3.0$	24.83 / 22.36 / 19.75	250.4 / 237.65 / 208.9
$w = 4.0$	26.22 / 23.84 / 21.48	**260.2** / 248.97 / 225.1

表 1: ImageNet 64x64 结果（$w = 0.0$ 指无引导模型）。

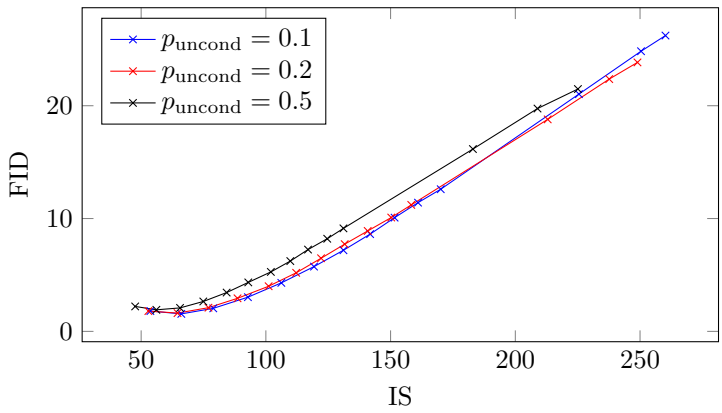


图 4: ImageNet 64x64 模型在不同引导强度下的 IS/FID 曲线。每条曲线对应一个具有特定无条件训练概率 p_{uncond} 的模型。配合 Table 1 阅读。

Dhariwal & Nichol 报告称容量相对较小的分类器即可满足有效的分类器引导采样，这与我们在无分类器引导模型上观察到的现象一致。

4.3 改变采样步数

采样步数 T 已知对扩散模型的样本质量有重大影响，因此本小节在 128×128 ImageNet 模型上研究改变 T 的影响。Table 2 和 Fig. 5 展示了在一组引导强度上改变 $T \in \{128, 256, 1024\}$ 的效果。正如预期，增大 T 会改善样本质量；对该模型而言，$T = 256$ 在样本质量与采样速度之间取得了较好的平衡。

注意，$T = 256$ 与 ADM-G (Dhariwal & Nichol, 2021) 所用采样步数大致相同，而我们的模型超过了 ADM-G。但需要注意，我们的方法在每个采样步中需要对去噪模型评估两次：一次评估条件得分 $\boldsymbol{\epsilon}_\theta(\mathbf{z}_\lambda, \mathbf{c})$，一次评估无条件得分 $\boldsymbol{\epsilon}_\theta(\mathbf{z}_\lambda)$。由于我们采用了与 ADM-G 相同的模型架构，就采样速度而言的公平比较应是我们 $T = 128$ 的设置，而该设置在 FID 分数上劣于 ADM-G。

Model	FID (↓)	IS (↑)
BigGAN-deep, max IS (Brock et al., 2019)	25	253
BigGAN-deep (Brock et al., 2019)	5.7	124.5
CDM (Ho et al., 2021)	3.52	128.8
LOGAN (Wu et al., 2019)	3.36	148.2
ADM-G (Dhariwal & Nichol, 2021)	2.97	-
Ours	$T = 128/256/1024$	
$w = 0.0$	8.11 / 7.27 / 7.22	81.46 / 82.45 / 81.54
$w = 0.1$	5.31 / 4.53 / 4.5	105.01 / 106.12 / 104.67
$w = 0.2$	3.7 / 3.03 / 3	130.79 / 132.54 / 130.09
$w = 0.3$	3.04 / **2.43** / **2.43**	156.09 / 158.47 / 156
$w = 0.4$	3.02 / 2.49 / 2.48	183.01 / 183.41 / 180.88
$w = 0.5$	3.43 / 2.98 / 2.96	206.94 / 207.98 / 204.31
$w = 0.6$	4.09 / 3.76 / 3.73	227.72 / 228.83 / 226.76
$w = 0.7$	4.96 / 4.67 / 4.69	247.92 / 249.25 / 247.89
$w = 0.8$	5.93 / 5.74 / 5.71	265.54 / 267.99 / 265.52
$w = 0.9$	6.89 / 6.8 / 6.81	280.19 / 283.41 / 281.14
$w = 1.0$	7.88 / 7.86 / 7.8	295.29 / 297.98 / 294.56
$w = 2.0$	15.9 / 15.93 / 15.75	378.56 / 377.37 / 373.18
$w = 3.0$	19.77 / 19.77 / 19.56	409.16 / 407.44 / 405.68
$w = 4.0$	21.55 / 21.53 / 21.45	**422.29** / 421.03 / 419.06

表 2: ImageNet 128x128 结果（$w = 0.0$ 指无引导模型）。

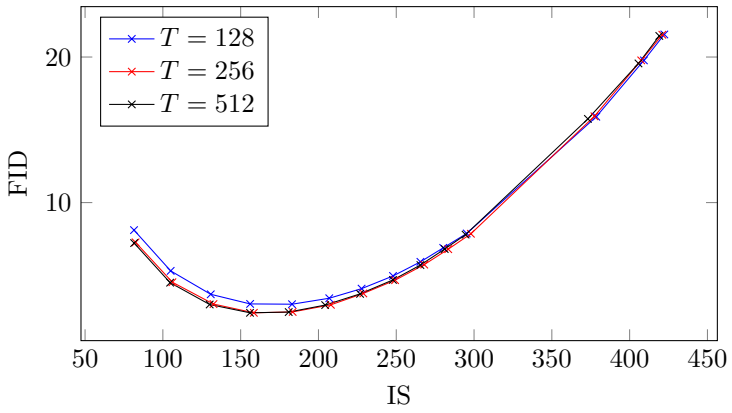


图 5: ImageNet 128x128 模型在不同引导强度下的 IS/FID 曲线。每条曲线对应不同的采样步数 T。配合 Table 2 阅读。

5 讨论

无分类器引导方法最实用的优点在于其极其简单：训练时只需改动一行代码——随机丢弃条件信息；采样时同样只需一行——将条件得分估计与无条件得分估计混合。相比之下，分类器引导会使训练流程复杂化，因为它需要额外训练一个分类器。该分类器必须在带噪声的 $\mathbf{z}_\lambda$ 上训练，因此无法直接套用标准的预训练分类器。

无分类器引导无需额外训练分类器，便能与分类器引导一样在 IS 与 FID 之间权衡，由此我们证明：引导可以由纯生成模型完成。此外，我们的扩散模型由无约束神经网络参数化，其得分估计未必构成保守向量场，这与分类器梯度不同 (Salimans & Ho, 2021)。因此，我们的无分类器引导采样器所遵循的步进方向与分类器梯度毫无相似之处，不能解释为针对分类器的基于梯度的对抗攻击；这进一步表明，借助一种不利用分类器梯度、不对图像分类器构成对抗的采样过程，纯生成模型同样能提升基于分类器的 IS 与 FID 指标。

我们还得到了关于引导如何起作用的直观解释：它在降低样本无条件似然的同时提高条件似然。无分类器引导通过一个负的得分项来降低无条件似然从而实现这一点，据我们所知这一思路此前尚未被探索，或许能在其他应用中派上用场。

此处所述的无分类器引导依赖训练一个无条件模型，但在某些情形下可以避免这一点。若类别分布已知且类别数较少，可利用 $\sum_{\mathbf{c}} p(\mathbf{x}|\mathbf{c})p(\mathbf{c}) = p(\mathbf{x})$ 这一关系，从条件得分得到无条件得分，而无需为无条件得分单独训练。当然，这样做需要与 $\mathbf{c}$ 取值个数相同的前向传播次数，对高维条件而言并不高效。

无分类器引导一个潜在的缺点是采样速度。通常分类器比生成模型更小更快，因此分类器引导采样可能快于无分类器引导，因为后者需要对扩散模型进行两次前向传播，一次算条件得分，一次算无条件得分。多次前向传播的必要性或许能通过调整网络架构、让条件信息在网络较后阶段才注入来缓解，但我们将这一探索留作未来工作。

最后，任何以牺牲多样性为代价提升样本保真度的引导方法，都必须面对“多样性下降是否可接受”这一问题。在已部署的模型中可能产生负面影响：在数据的某些部分相对其余部分表达不足的应用里，维持样本多样性十分重要。如何在保持样本多样性的同时提升样本质量，是一个值得探索的未来方向。

6 结论

我们提出了无分类器引导，一种在扩散模型中提升样本质量、降低样本多样性的方法。无分类器引导可看作不带分类器的分类器引导，其有效性表明：纯生成扩散模型能够在完全避开分类器梯度的同时，最大化基于分类器的样本质量指标。我们期待在更广泛的设定与数据模态下进一步探索无分类器引导。

7 致谢

我们感谢 Ben Poole 与 Mohammad Norouzi 的讨论。

参考文献

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.

Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. WaveGrad: Estimating gradients for waveform generation. *International Conference on Learning Representations*, 2021.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat GANs on image synthesis. *arXiv preprint arXiv:2105.05233*, 2021.

Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, pp. 529–536, 2004.

Peter Grünwald and John Langford. Suboptimal behavior of bayes and mdl in classification under misspecification. *Machine Learning*, 66(2-3):119–149, 2007.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, pp. 6626–6637, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pp. 6840–6851, 2020.

Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *arXiv preprint arXiv:2106.15282*, 2021.

Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

Diederik P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, pp. 10215–10224, 2018.

Diederik P Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *arXiv preprint arXiv:2107.00630*, 2021.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. DiffWave: A Versatile Diffusion Model for Audio Synthesis. *International Conference on Learning Representations*, 2021.

Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. *International Conference on Machine Learning*, 2021.

Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with VQ-VAE-2. In *Advances in Neural Information Processing Systems*, pp. 14837–14847, 2019.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Tim Salimans and Jonathan Ho. Should EBMs model the energy or the score? In *Energy Based Models Workshop-ICLR 2021*, 2021.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*, pp. 2234–2242, 2016.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265, 2015.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, pp. 11895–11907, 2019.

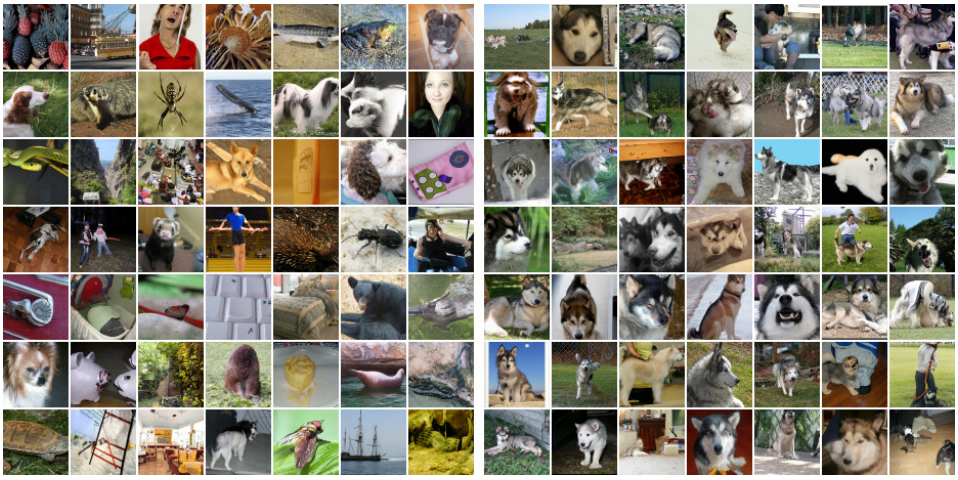
Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *arXiv e-prints*, pp. arXiv–2101, 2021a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *International Conference on Learning Representations*, 2021b.

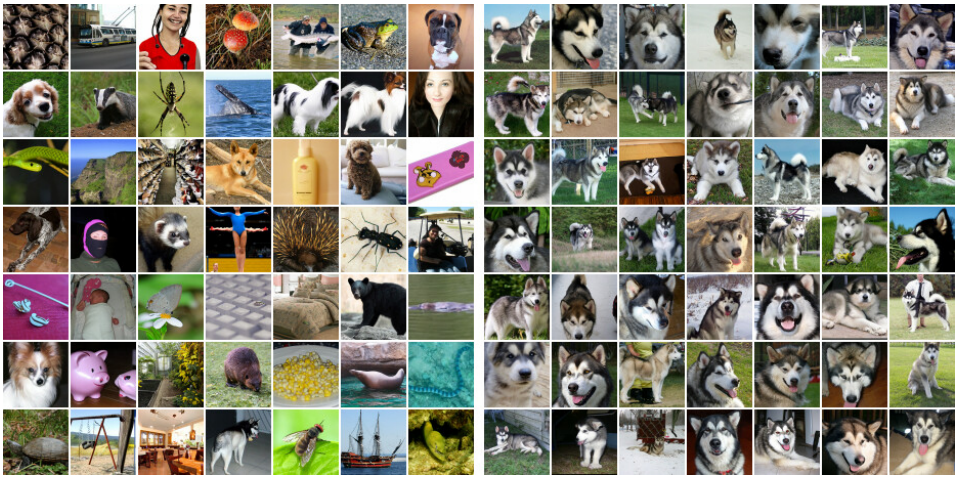
Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

Yan Wu, Jeff Donahue, David Balduzzi, Karen Simonyan, and Timothy Lillicrap. LOGAN: Latent optimisation for generative adversarial networks. *arXiv preprint arXiv:1912.00953*, 2019.

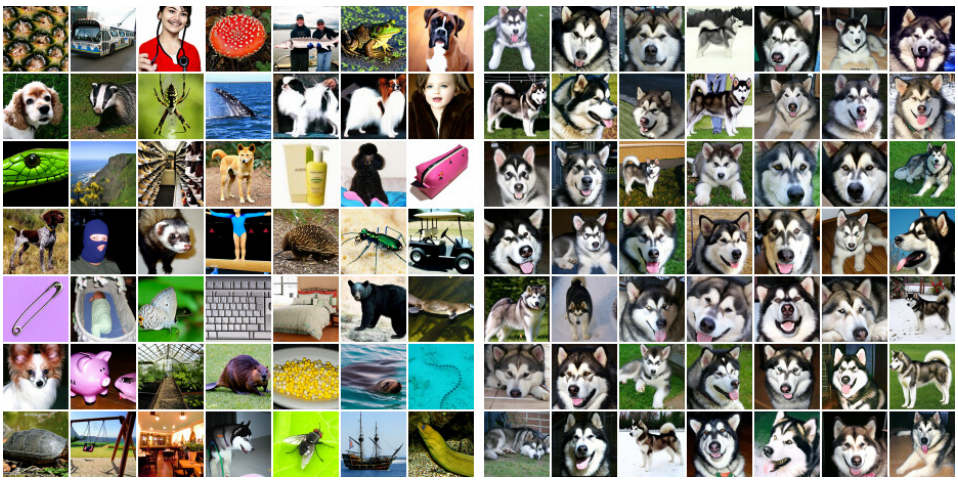
A 样本



(a) 无引导条件采样：FID=1.80，IS=53.71

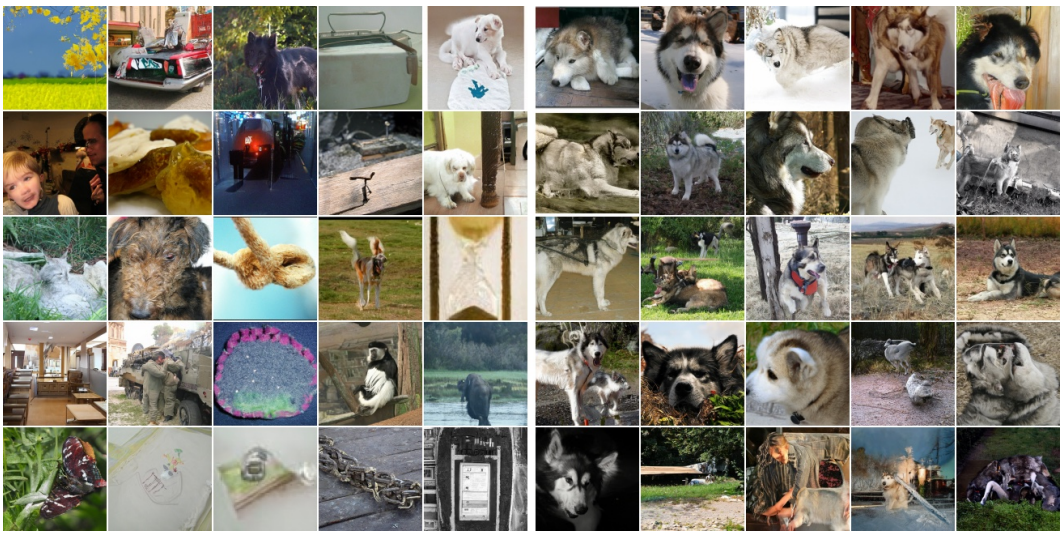


(b) 无分类器引导，$w = 1.0$：FID=12.6，IS=170.1

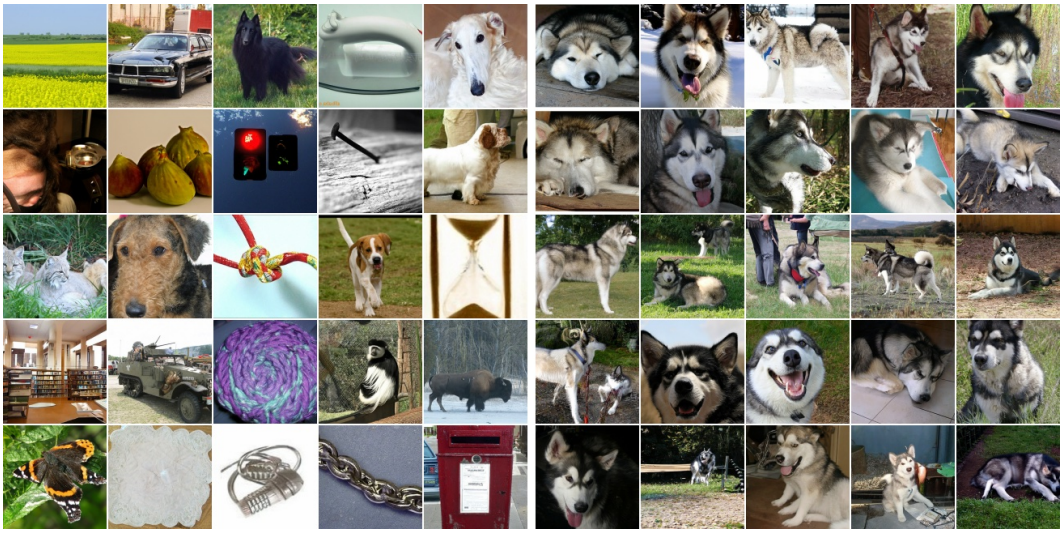


(c) 无分类器引导，$w = 3.0$：FID=24.83，IS=250.4

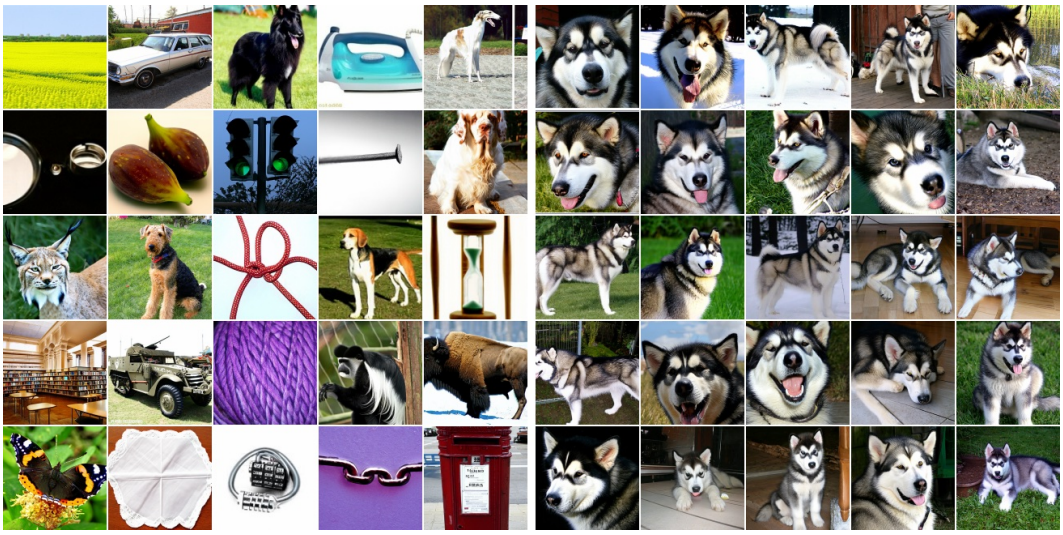
图 6: ImageNet 64x64 上的无分类器引导。左：随机类别。右：单一类别（malamute）。各子图采样使用相同的随机种子。



(a) 无引导条件采样：FID=7.27，IS=82.45



(b) 无分类器引导，$w = 1.0$：FID=7.86，IS=297.98



(c) 无分类器引导，$w = 4.0$：FID=21.53，IS=421.03

图 7: ImageNet 128x128 上的无分类器引导。左：随机类别。右：单一类别（malamute）。各子图采样使用相同的随机种子。

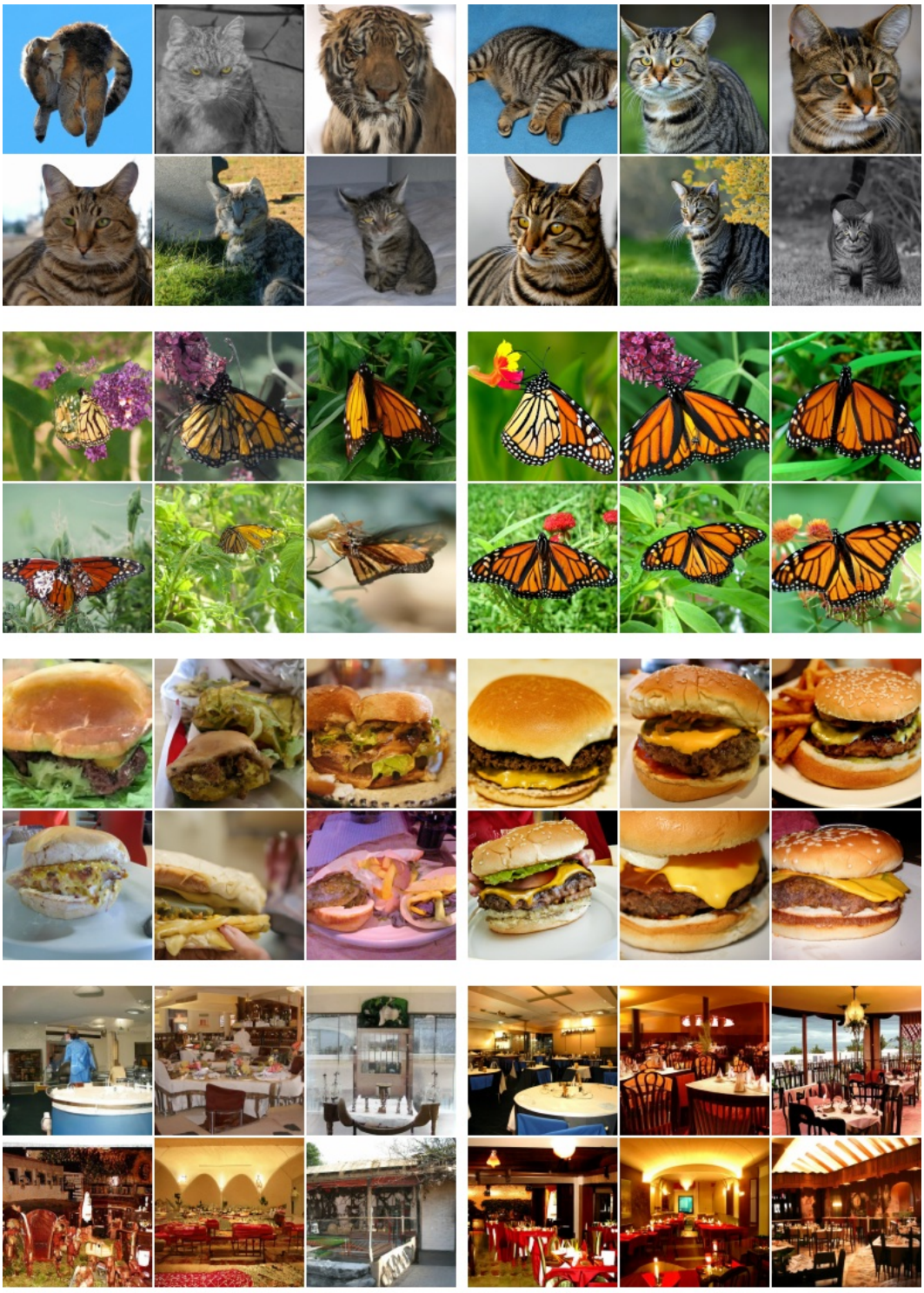


图 8: 128x128 ImageNet 上无分类器引导的更多示例。左：无引导样本；右：$w = 3.0$ 的无分类器引导样本。