缩放整流流量变压器以实现高分辨率图像合成

帕特里克·埃塞尔 * 苏米斯·库拉尔 安德烈亚斯·布拉特曼 拉希姆·恩特扎里 乔纳斯·穆勒 哈里·赛尼 亚姆·莱维
多米尼克·洛伦茨 阿克塞尔·绍尔 弗雷德里克·博塞尔 达斯汀·波德尔 蒂姆·多克霍恩 锡安 英语
凯尔·莱西 亚历克斯·古德温 扬尼克·马雷克 罗宾·隆巴赫 *
稳定性人工智能

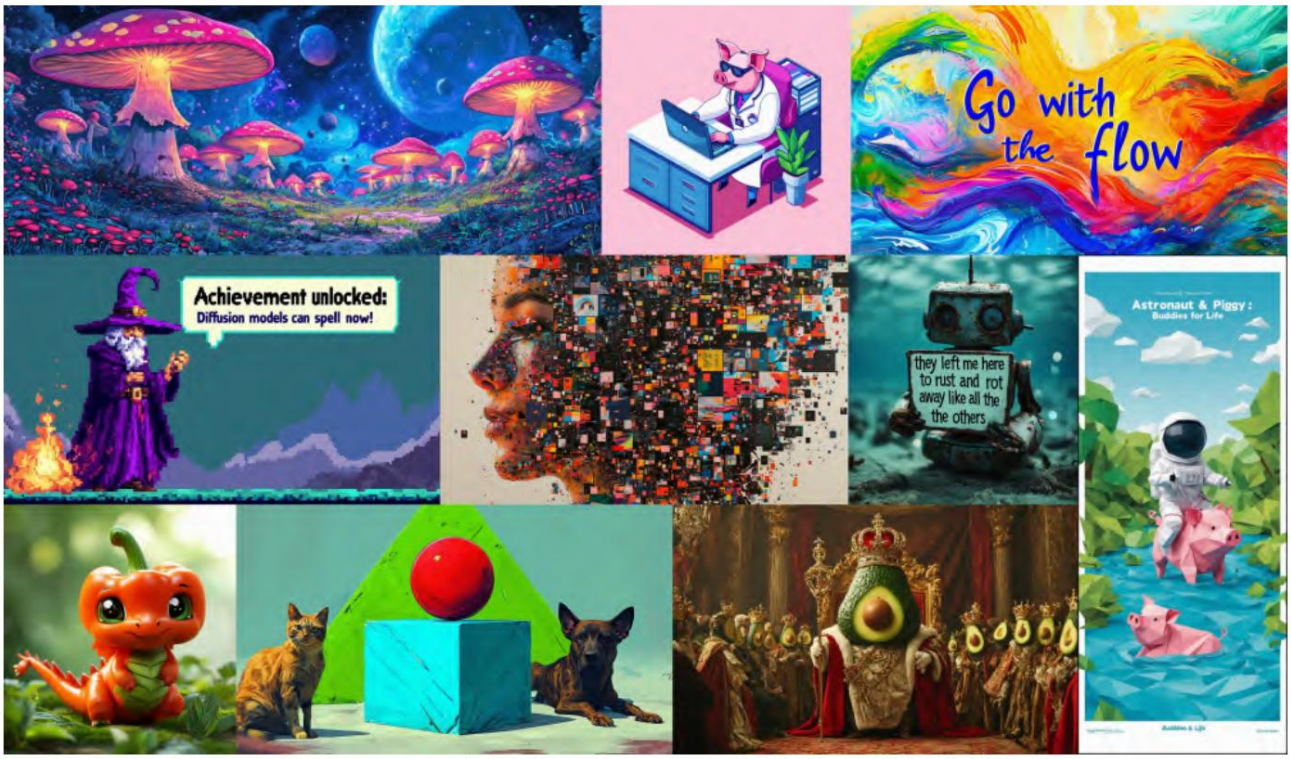


图 1.来自我们的 8B 整流流模型的高分辨率样本,展示了其排版、精确提示跟随方面的功能和空间推理、对细节的关注以及各种风格的高图像质量。

抽象的

扩散模型通过将数据的前向路径反转为噪声来创建数据。
已经成为一种强大的生成模型
高维感知数据技术
例如图像和视频。整流流是一种最新的生成模型公式,它将
数据和噪声成一条直线。尽管它更好
理论特性和概念简单性,它
尚未被明确确立为标准实践。在这项工作中,我们改进了现有的噪声采样技术,通过将修正流模型偏向于感知相关的模型来训练它们

秤。通过大规模研究,我们证明
阐述该方法的优越性能
与已建立的扩散配方相比
用于高分辨率文本到图像的合成。此外,我们提出了一种新颖的基于变压器的
用于文本到图像生成的架构,使用
两种模式的权重分开,并实现信息之间的双向流动

图像和文本标记,提高文本理解、排版和人类偏好评级。
我们证明该架构遵循可预测的缩放趋势,并将较低的验证损失与改进的文本到图像合成相关联,如下所示
通过各种指标和人类评估来衡量。我们最大的模型优于最先进的模型,我们将进行实验

数据、代码和模型权重公开。

*平等贡献。 <first.last>@stability.ai。

一、简介

扩散模型从噪声中创建数据（Song 等人， 2020）。
它们被训练将数据的前向路径反转为随机噪声，因此，结合神经网络的近似和泛化特性，可以用来生成训练数据中不存在但遵循训练分布的新数据点数据（Sohl-Dickstein 等人， 2015 年； Song 和 Ermon， 2020 年）。

这种生成建模技术已被证明对于对图像等高维感知数据进行建模非常有效（Ho et al., 2020）。近年来，扩散模型已成为从自然语言输入生成高分辨率图像和视频的事实上的方法，具有令人印象深刻的泛化能力（Saharia等人， 2022b； Ramesh 等人， 2022； Rombach 等人， 2022 ）； Podell 等人， 2023； Dai 等人， 2023； Esser等人， 2023； Blattmann 等人， 2023b； Betker 等人， 2023； Blattmann 等人， 2023a； Singer 等人， 2022） 。由于其迭代性质和相关的计算成本，以及推理过程中的较长采样时间，对这些模型进行更有效训练和/或更快速采样的公式的研究有所增加（Karras 等人， 2023 年； Liu 等人，2023 年）。 ， 2022）。

虽然指定从数据到噪声的前向路径可以实现高效的训练，但它也提出了选择哪条路径的问题。这种选择可能对采样产生重要影响。例如，无法消除数据中所有噪声的前向过程可能会导致训练和测试分布的差异，并导致诸如灰度图像样本之类的伪影（Lin 等人， 2024）。重要的是，前向过程的选择也会影响学习到的后向过程，从而影响采样效率。虽然弯曲路径需要许多积分步骤来模拟该过程，但直线路径可以用单个步骤进行模拟，并且不易出现误差累积。由于每个步骤都对应于神经网络的评估，因此这对采样速度有直接影响。

前向路径的一个特殊选择是所谓的整流流（Liu et al., 2022; Albergo & Vanden-Eijnden, 2022; Lipman et al., 2023），它将数据和噪声连接在一条直线上。尽管该模型类具有更好的理论特性，但尚未在实践中得到决定性的确立。到目前为止，一些优势已经在中小型实验中得到了实证证明（ Ma et al., 2024），但这些大多局限于类条件模型。在这项工作中，我们通过在修正流模型中引入噪声尺度的重新加权来改变这一点，类似于噪声预测扩散模型（Ho et al., 2020）。通过大规模研究，我们将新配方与现有扩散配方进行比较并证明其优点。

我们展示了广泛使用的文本到图像合成方法，其中直接输入固定的文本表示到模型中（例如，通过交叉注意力（Vaswani 等人， 2017； Rombach 等人， 2022））并不理想，并且提出了一种新的架构，该架构结合了图像和文本标记的可学习流，这使得它们之间的双向信息流。我们将其与我们的改进了整流流公式并研究了其可扩展性。我们展示了验证损失的可预测缩放趋势，并表明较低的验证损失与改进的自动和人工评估密切相关。

我们最大的模型优于最先进的开放模型，例如 SDXL （Podell 等人， 2023）、 SDXL-Turbo （Sauer等人， 2023）、 Pixart-α （Chen 等人， 2023)和封闭模型-源模型，例如 DALL-E 3 (Betker et al., 2023) ，均用于即时理解和人类偏好评级的定量评估(Ghosh et al., 2023) 。

我们工作的核心贡献是：(i)我们对不同的扩散模型和整流流公式进行了大规模、系统的研究，以确定最佳设置。

为此，我们为整流流模型引入了新的噪声采样器，与之前已知的采样器相比，其性能得到了提高。 (ii) 我们设计了一种新颖的、可扩展的文本到图像合成架构，允许网络内文本和图像令牌流之间的双向混合。我们展示了它与 UViT (Hoogeboom et al., 2023)和 DiT (Peebles & Xie, 2023)等已建立的骨干网相比的优势。最后，我们 (iii) 对我们的模型进行缩放研究，并证明它遵循可预测的缩放趋势。我们表明，较低的验证损失与通过 T2I-CompBench （Huang 等人， 2023）、 GenEval （Ghosh 等人， 2023)和人类评分等指标评估的文本到图像性能的提高密切相关。我们公开结果、代码和模型权重。

2. 流程的无模拟训练

我们考虑使用常微分方程（ODE)定义噪声分布p1的样本x1到数据分布p0的样本x0之间的映射的生成模型，

$$dyt = v\theta(yt, t)\, dt \quad , \tag{1}$$

其中速度v由神经网络的权重θ参数化。 Chen 等人之前的工作。 (2018)建议通过可微 ODE求解器直接求解方程(1) 。然而，这个过程的计算成本很高，特别是对于参数化 vθ(yt, t)的大型网络架构。更有效的替代方法是直接回归向量场ut ，生成p0和p1之间的概率路径。为了构建这样的ut，我们定义一个前向过程，对应于p0和p1 = N (0, 1)之间的概率路径pt ，如下

$$zt = atx0 + bt其中 \quad N(0, I) 。 \tag{2}$$

对于a0 = 1、 b0 = 0、 a1 = 0 和b1 = 1,边际，

$$pt(zt) = E \quad N(0,I)pt(zt|) , \tag{3}$$

与数据和噪声分布一致。

将zt、 x0与duce ψt和ut之间的关系表示为 , 我们介绍-

$$\psi t(\cdot|) : x0 \to atx0 + bt \tag{4}$$

$$ut(z|) := \psi \quad {}_t \, (\psi_t^{-1}(z|)|) \tag{5}$$

zt可以写为初始值z0 = x0的 ODE z的解,因此 ut(·|)生成 $_t$ = ut(zt|),由于 pt(·|)。值得注意的是,我们可以使用条件向量场ut(·|)构造一个边缘向量场ut , 它生成边缘概率路径pt (Lipman et al., 2023) （参见B.1） ：

$$ut(z) = E \quad N(0,I)ut(z|) \, pt(z) \quad \frac{pt(z|)}{} \tag{6}$$

使用流量匹配目标回归ut时

$$LFM = Et,pt(z) \, ||v\theta(z, t) - ut(z)||2 \, 2 。 \tag{7}$$

由于方程 6 中的边缘化,条件流匹配（参见 B.1） ,直接处理是很棘手的，

$$LCFM = Et,pt(z|),p() \, ||v\theta(z, t) - ut(z|)||2 \, 2 , \tag{8}$$

利用条件向量场ut(z|)提供了一个等效但易于处理的目标。

为了将损失转换为显式形式,我们将z − bt ψ和ψ插入到(5)中

$$_t(x0|) = a \, tx0 + b \quad {}_t \quad {}_t^{-1}(z|) = \quad \frac{}{}$$

$$zt \quad = ut(zt|) = \quad \frac{在}{在} zt - bt(\quad \frac{在}{在} - \frac{乙_t}{BT}) 。 \tag{9}$$

现在,考虑信噪比λt := log $\frac{a_t^2}{b_t^2}$。和 $\lambda_t = 2(\frac{在}{在} - \frac{乙_t}{BT})$,我们可以将方程（9)重写为

$$ut(zt|) = \quad \frac{a_t}{zt - 在} \quad \frac{bt}{2}\lambda_t \tag{10}$$

接下来,我们使用方程（10)重新参数化方程（8)作为噪声预测目标：

$$LCFM = Et,pt(z|),p() \, ||v\theta(z, t) - \quad \frac{a_t}{\lambda z + 2 \, at\frac{bt}{}} \quad {}_t \quad ||2 \, 2 \tag{11}$$

$$= Et,pt(z|),p() - \lambda \, 2 \quad \frac{BT}{} \quad {}_t \quad ||\theta(z, t) - ||2 \, 2 \, (12)$$

我们定义θ := $\frac{-2\lambda}{_t BT}$ (vθ- $\frac{在}{在}$ z) 。

请注意,当引入时间相关加权时,上述目标的最优值不会改变。因此，

人们可以推导出各种加权损失函数,这些函数为所需的解决方案提供信号,但可能会影响优化轨迹。为了对不同方法（包括经典的扩散公式)进行统一分析,我们可以将目标写成以下形式（遵循Kingma&Gao （2023）） ：

$$Lw(x0) = - \quad \frac{1}{2} Et \quad U(t), \quad N(0,I) \, wt\lambda \, t\theta(zt, t) - \quad {}^2 \, ,$$

其中wt = - $\frac{1}{2}\lambda_t z_t^2$ 对应于LCFM。

3. 流动轨迹

在这项工作中,我们考虑了上述形式主义的不同变体,我们将在下面简要描述。

整流流整流流（RFs） （Liu et al., 2022; Albergo & Vanden-Eijnden, 2022; Lipman et al., 2023)将前向过程定义为数据分布和标准正态分布之间的直线路径,即

$$zt = (1 - t)x0 + t \quad , \tag{13}$$

并使用LCFM ,然后对应于w 网络输出直接参数化速度vθ。 $\frac{射频}{t} = \frac{}{t\,1-t}$。

EDM EDM (Karras et al., 2022)使用以下形式的前向过程

$$zt = x0 + bt \tag{14}$$

Gau, 2023) bt = exp F) − 1其中F N是正态分布的分位数函数。注意这个选择 $^{-1}$ N (t|Pm, P2其中(Kingma &S

化与均值Pm和方差P 结果 ^2_s

$$\lambda t \quad N (-2Pm,(2Ps) \quad {}^2) \text{ 对于 } t \quad U(0, 1) \tag{15}$$

网络通过F 预测进行参数化（Kingma&Gao， 2023； Karras 等人， 2022） ,并且损失可以写为LwEDM：

$$_t$$

$$\frac{电火花加}{工重量} = N \, (\lambda t| - 2Pm,(2Ps) \quad {}^2)(e^{-\lambda t} + 0.5 \quad {}^2) \tag{16}$$

Cosine (Nichol & Dhariwal, 2021)提出了以下形式的前向过程

$$zt = \cos\pi \, t \, x\theta_2 + \sin\pi \quad \frac{}{2}t \quad 。 \tag{17}$$

与参数化和损失相结合,这对应于权重wt = sech(λt/2)。当与v 预测损失相结合时（Kingma &Gao， 2023） ， − λt/2权重由下式给出： wt = e

。

(LDM-)线性LDM （Rombach等人， 2022)使用DDPM时间表的修改（ Ho等人， 2020）。两者都是方差保留计划,即bt = 1 −离散时间步长t = 0的细化，...。 。 。 ,扩散系数βt为at = (ts=0(1 − βs)) 2 t, 并去 T − 1表示

对于给定的边界值β0和βT -1， βt = β0 + t T−1 (βT −1 − β0)且 LDM 使用βt = β0 + t T−1 (βT −1 − β0) 2 。

DDPM用途 $\frac{d}{dt}f_{模式}^{-1}(t)$。 如图 11 所示,尺度参数控制采样过程中中点（正s)或端点（负s)的优先程度。该公式还包括s = 0时的均匀加权πmode(t; s = 0) = U(t) ,该公式已在之前的整流流工作中广泛使用（Liu et al., 2022； Ma et al., 2024） ）。

3.1.适用于 RF 模型的定制 SNR 采样器

RF 损失在[0, 1] 中的所有时间步上均匀地训练速度vθ 。然而,直观上,对于[0, 1]中间的t,所得到的速度预测目标− x0更困难,因为对于t = 0,最佳预测是p1的平均值,而对于t = 1 ,最佳预测是 p1 的平均值。预测是 p0 的平均值。一般来说,将t上的分布从常用的均匀分布U(t)更改为密度为π(t)的分布相当于加权损失Lwπ

$$_{重量}{}^{\pi} = \frac{t}{1-t}\pi(t) \tag{18}$$

因此,我们的目标是通过更频繁地采样中间时间步来赋予它们更多的权重。接下来，我们描述用于训练模型的时间步密度 π(t)。

Logit 正态采样对中间步骤给予更多重视的分布的一种选择是 Logit 正态分布（Atchison & Shen， 1980）。它的密度，

$$m, s) = \exp -s\sqrt{2\pi\, t(1-t)^{2s^2}}\,\frac{1}{1}\qquad \frac{(\text{logit}(t)-m)\,\pi \ln(t;}{}^{2}, \tag{19}$$

其中logit(t) = log具有位置参数t,m 和尺度参数s。位置参数使我们能够将训练时间步长偏向数据p0 （负m)或噪声p1 （正m）。如图11 所示,尺度参数控制分布的宽度。

在实践中,我们从正态分布u N (u; m, s)中对随机变量u进行采样,并通过标准逻辑函数将其映射。

重尾模式采样Logit-正态密度总是在端点0和1 处消失。为了研究这是否对性能产生不利影响,我们还使用了在[0, 1上具有严格正密度的时间步采样分布]。对于尺度参数s,我们定义

$$\text{fmode}(u; s) = 1 - u - s\cdot\cos^2\!\left(\frac{\pi}{2}u - 1 + u\right)。 \tag{20}$$

对于−1 ⩽ s ⩽该函数是单调的,我们可以使用它从隐含密度πmode(t; s) =中进行采样

CosMap最后,我们还考虑RF 设置中第3节的余弦时间表（Nichol & Dhariwal， 2021） 。

特别是,我们正在寻找映射f : u → f(u) = t, u ∈ [0, 1],使得 log-snr 与余弦cos(π u)时间表相匹配： 2 log f(你） 。求解f,我们u)

$$\frac{2}{\text{正弦}(\frac{\pi}{2}u)}\ 2\log\ \frac{1-f(u)=}{}$$

得到 u U(u)

$$t = f(u) = 1 - \tan\left(\pi\,\frac{1}{u) + 1,\ 2}\right) \tag{21}$$

从中我们得到密度

$$\pi\text{CosMap}(t) = \frac{d}{dt}f^{-1}(t) = \pi\,\frac{2}{-2\pi t + 2\pi t2}。 \tag{22}$$

4. 文本到图像架构对于图像的文本条件采样,我们的模型必须考虑文本和图像这两种模式。我们使用预训练的模型来导出合适的表示,然后描述我们的扩散主干的架构。图 2 对此进行了概述。

我们的一般设置遵循 LDM （Rombach 等人， 2022）,用于在预训练自动编码器的潜在空间中训练文本到图像模型。与将图像编码为潜在表示类似,我们也遵循以前的方法（Saharia 等人， 2022b； Balaji 等人， 2022)并使用预训练的冻结文本模型对文本条件c进行编码。详细信息请参见附录B.2。

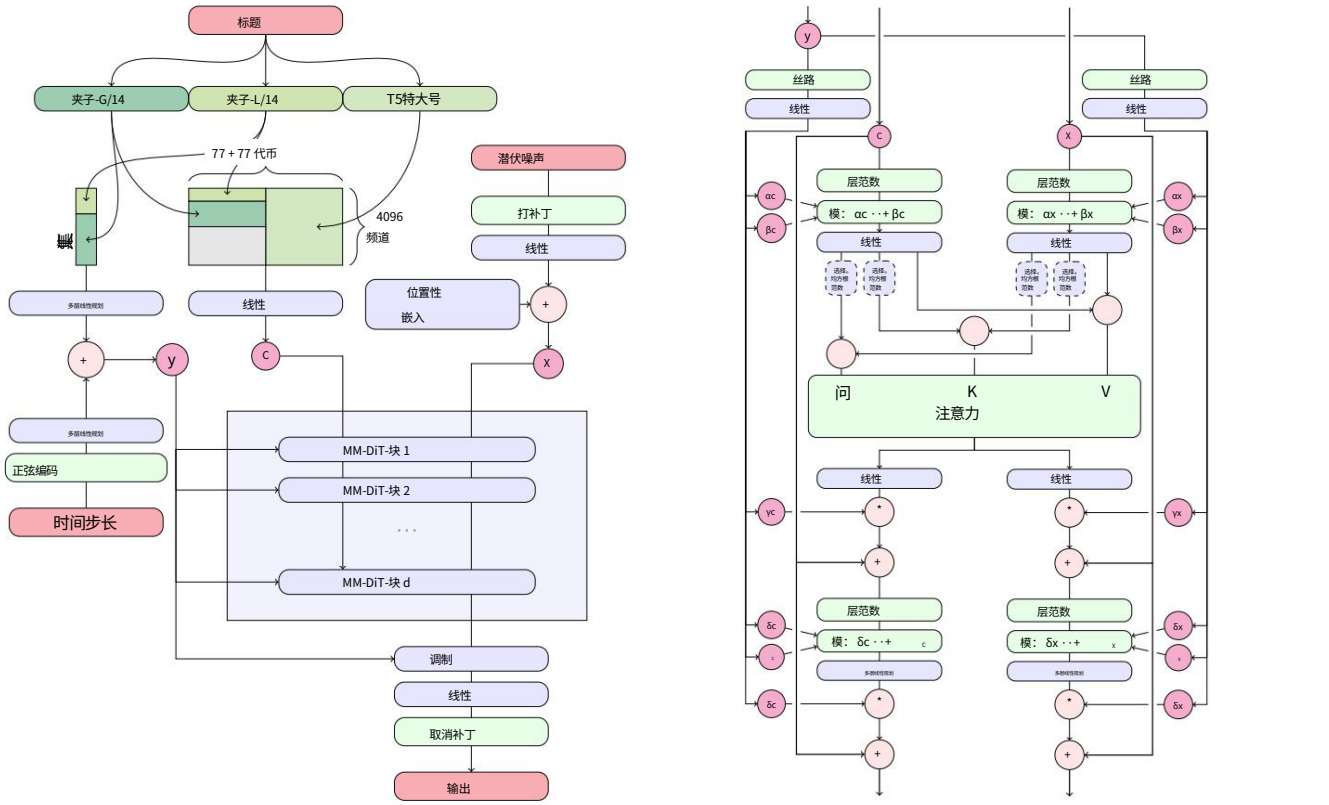
多模态扩散主干我们的架构建立在 DiT (Peebles & Xie, 2023)架构之上。 DiT 仅考虑类条件图像生成,并使用调制机制根据扩散过程的时间步长和类标签来调节网络。

类似地,我们使用时间步t和cvec的嵌入作为调制机制的输入。然而,由于池化文本表示仅保留有关文本输入的粗粒度信息（Podell 等人， 2023）,因此网络还需要来自序列表示的信息

cctxt。

我们构建一个由文本和图像输入的嵌入组成的序列。具体来说,我们添加位置编码并将潜在像素代表的2 × 2块展平- h×w×c表示x ∈ R

到长度· w的补丁编码序列。将这个 patch 编码和文本编码cctxt嵌入到一个共同的维度之后,我们



(a) 所有组成部分的概述。

(b) 一个 MM-DiT 块

图 2.我们的模型架构。连接由 和逐元素乘法 * 表示。可以添加Q和K的 RMS-Norm以稳定训练运行。最佳观看放大。

连接两个序列。然后，我们遵循 DiT 并应用一系列调制注意力和 MLP。

由于文本和图像嵌入在概念上完全不同，因此我们对这两种模式使用两组独立的权重。如图2b 所示，这相当于每种模态都有两个独立的转换器，但是将两种模态的序列连接起来进行注意力操作，这样两种表示都可以在自己的空间中工作，同时考虑另一种表示。

对于我们的缩放实验，我们通过将隐藏大小设置为64·d （在 MLP 块中扩展到4·64·d个通道），根据模型深度d（即注意块的数量）参数化模型的大小。 ，注意力头的数量等于 d。

5. 实验

5.1.改善整流流程

我们的目标是了解公式1中哪种无模拟训练归一化流的方法是最有效的。为了能够比较不同方法，我们控制优化算法、模型架构、数据集和采样器。在

此外，不同方法的损失是不可比的，也不一定与输出样本的质量相关；因此，我们需要能够对方法进行比较的评估指标。我们在 ImageNet（Russakovsky 等人，2014)和 CC12M（Changpinyo 等人，2021)上训练模型，并在训练过程中使用验证损失、 CLIP 分数（Radford 等人)评估模型的训练和 EMA 权重., 2021; Hessel et al., 2021)和不同采样器设置（不同引导尺度和采样步骤)下的FID (Heusel et al., 2017) 。我们按照(Sauer et al., 2021)的建议计算 CLIP 特征的 FID 。所有指标均在 COCO-2014 验证分割上进行评估（Lin 等人， 2014）。附录 B.3 中提供了有关训练和采样超参数的完整详细信息。

5.1.1.结果

我们在两个数据集上训练了 61 种不同的公式。我们包括第 3 节中的以下变体：

- 线性（eps/线性、 v/线性)和余弦（eps/cos、 v/cos)计划的- 和v 预测损失。
- πmode(t; s) (rf/mode(s))的 RF 损耗， s的7 个值统一在-1和1.75之间选择，并且