用于高分辨率图像合成的比例整流变压器

Patrick Esser [*] **Sumith Kulal Andreas Blattmann Rahim Entezari Jonas Mu¨ ller Harry Saini Yam Levi Dominik Lorenz Axel Sauer Frederic Boesel Dustin Podell Tim Dockhorn Zion English**
罗宾-隆巴赫 [*]
稳定性人工智能

*图 1.*我们的 8B 整流模型的高分辨率样本，展示了它在排版、精确的提示跟踪和空间推理、对精细细节的关注以及各种风格的高图像质量方面的能力。

摘要

扩散模型通过反转数据向噪声的前向路径，从噪声中创建数据，已成为图像和视频等高维感知数据的强大生成模型技术。整流是一种将数据和噪声以直线连接起来的再生成模型表述。尽管它具有较好的理论特性和概念上的简洁性，但尚未被明确确立为标准实践。在这项工作中，我们改进了用于训练整流模型的现有噪声采样技术，使其偏向于与感知相关的尺度。通过大规模研究，我们证明了

在高分辨率文本到图像的合成方面，与已有的扩散公式相比，我们的方法具有更优越的性能。从传统意义上讲，我们提出了一种基于变压器的文本到图像生成新架构，该架构对两种模态使用不同的权重，并使图像和文本标记之间的信息流能够双向流动，从而改善了文本理解、排版和人类偏好评级。我们证明，这种架构遵循了可预设的扩展趋势，并通过各种指标和人工评估，将较低的估价损失与改进的文本到图像合成联系起来。我们的最大模型优于最先进的模型。Stability AI 正在考虑公开实验数据、代码和模型权重。

(< first.last> @stability.ai.

第41 (st) 届国际机器学习大会论文集，奥地利维也纳。PMLR 235, 2024.

1. 简介

扩散模型从噪声中创建数据（Song 等人，2020 年）。这些模型经过训练，可将数据的前向路径反转为随机噪声，因此，结合神经网络的逼近和泛化特性，可用于生成训练数据中不存在但遵循训练数据分布的新数据点（Sohl-Dickstein 等人，2015 年；Song & Ermon，2020 年）。事实证明，这种生成建模技术对图像等高维感知数据的建模非常有效（Ho 等人，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方法，其泛化能力令人印象深刻（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）。由于这些模型的迭代性质和相关计算成本，以及推理过程中的长时间采样，对这些模型进行更高效训练和/或更快速采样的研究有所增加（Karras 等，2023 年；Liu 等，2022 年）。

虽然指定一条从数据到噪声的前向路径可以提高训练效率，但同时也提出了选择哪条路径的问题。这种选择会对采样产生重要影响。例如，如果前向过程无法去除数据中的 所有噪声，就会导致训练和测试分布不一致，并产生灰度图像样本等伪影（Lin 等人，2024 年）。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。曲线路径需要多个积分步骤来模拟，而直线路径只需一个步骤即可模拟，且不易累积误差。由于每一步都对应着对神经网络的评估，这对采样速度有直接影响。

前向路径的一种特殊选择是所谓的 "*简化流*"（Liu 等，2022；Albergo & Vanden-Eijnden，2022；Lipman 等，2023），它在一条直线上连接数据和噪声。虽然这类模型具有较好的理论特性，但在实践中还没有得到决定性的应用。迄今为止，一些优势已在中小型实验中得到了验证（Ma 等人，2024 年），但这些优势大多局限于类条件模型。在这项工作中，我们通过在整流模型中引入噪声尺度的重新加权来改变这种情况，类似于噪声预测扩散模型（Ho 等人，2020 年）。通过大规模研究，我们将新方法与现有的扩散方法进行了比较，并证明了其优势。

我们发现，在文本到图像的合成中，广泛使用的方法是直接输入固定的文本表示。

因此，我们提出了一种新的架构，将图像和文本标记的可学习流结合在一起，从而实现它们之间的双向信息流。我们将其与改进的整流公式相结合，并对其可扩展性进行了研究。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自动和人工评估的改进密切相关。

我们的最大模型在及时理解的量化评估（Ghosh 等人，2023 年）和人类偏好评级方面都优于 *SDXL*（Podell 等人，2023 年）、SDXL-*Turbo*（Sauer 等人，2023 年）、*Pixart-α*（Chen 等人，2023 年）等最先进的开放模型和 DALL-E 3（Betker 等人，2023 年）等封闭源模型。

我们工作的核心贡献是(i) 我们对不同的扩散模型和整流公式进行了大规模的系统研究，以确定最佳设置。为此，我们为整流模型引入了新的噪声采样器，与之前已知的采样器相比，性能有所提高。(ii) 我们为文本到图像的合成设计了一种新颖、可扩展的架构，允许网络内文本和图像标记流之间的双向混合。我们展示了它与 UViT（Hoogeboom 等人，2023 年）和 DiT（Pee- bles & Xie，2023 年）等已建立的骨架相比的优势。最后，我们（iii）对我们的模型进行了缩放研究，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 年）、GenEval（Ghosh 等人，2023 年）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. 流量的无模拟训练

我们考虑的生成模型是通过普通差分方程（ODE）定义噪声分布 p_1 的样本 x_1与数据分布 p_0的样本 x_0之间的映射、

$$dy_t = v_{(\Theta)}(y_t, t)\, dt \text{ 、} \tag{1}$$

其中，速度 v 由神经网络的 权重 Θ 参数化。Chen 等人（2018 年）之前的工作建议通过可微 ODE 求解器直接求解方程 (1)。然而，这一过程的计算成本很高，尤其是对于参数化 $v_{(\Theta)}(y_t, t)$ *的*大型网络架构而言。更有效的方法是直接回归一个向量场 u_t，生成 p_0和 p_1之间的概率路径。为了构建这样的 u_t，我们定义一个前向过程，对应于 $p_(0)$和 $p_1 = N(0, 1)$ 之间的概率路径 p_t，即

$$z_t = a_t x_0 + b_{(t)}\epsilon \quad \text{其中 } \epsilon \sim N(0, I) \text{ 。} \tag{2}$$

对于 $a_0 = 1$，$b_0 = 0$，$a_1 = 0$ 和 $b_1 = 1$，边际值、

$$p_{(t)}(z_t) = \mathsf{E}_{\epsilon \sim N(0,I)}\left[p_{(t)}(z_t \mid \epsilon)\right] 、 \tag{3}$$

与数据和噪声分布一致。

为了表达 z_t、x_0和 ϵ之间的关系，我们引入了 ψ_t和 u_t的关系。

$$\psi_{(t)}(\cdot \mid \epsilon) : x_{(0)} \mapsto a_{(t)} x_{(0)} + b_{(t)} \epsilon \tag{4}$$

$$u_t(z \mid \epsilon) := \dot{\psi}_{(t)}\left(\psi_t^{-1}(z \mid \epsilon) \mid \epsilon\right) \tag{5}$$

由于 z_t可以写成 ODE $\dot{z}_t = u_{(t)}(z_t)$ ϵ的解，初始值为 $z_0 = x_{(0)}$，因此 $u_t(\epsilon)$会产生 $p_{(t)}(\epsilon)$。显然，我们可以利用条件向量场 $\dot{u}_t(\cdot \mid \epsilon)$构造一个边际向量场 $u_{(t)}$，它生成边际概率路径 $p_{(t)}$（Lipman 等人，2023 年）（见 B.1）：

$$u_t(z) = \mathsf{E}_{\epsilon \sim N(0,I)}\left[\frac{u_t(z \mid \epsilon)\, p(z \mid \epsilon)}{p_{(t)}(z)}\right] \tag{6}$$

而用*流量匹配*目标回归 u_t

$$L_{FM} = \mathsf{E}_{(t,)p_{(\cdot)}(z)}\left\| v_{(\Theta)}(z, t) - u_{(\cdot)}(z) \right\|_2^2 \tag{7}$$

由于等式中的边际化，直接计算是难以实现的。
第 6 章，*条件流匹配*（见 B.1）、

$$L_{CFM} = \mathsf{E}_{(t,p)(\cdot)(z)\mid(\epsilon)(),p()(\epsilon)()}\left\| v_{(\Theta)}(z, t) - u_{(t)}(z \mid \epsilon) \right\|_2 、 \tag{8}$$

与条件矢量场 $u_{(t)}(z \mid \epsilon)$ 的关系提供了一个等价而又简单的目标。

为了将损失转换为明确的形式，我们插入 $\psi_t(x_{(0)} \mid \epsilon) = a_t x_0 + b_t \epsilon$ 和 $\psi_t^{-1}(z \mid \epsilon) = \frac{z - b_t \epsilon}{a_t}$ 插入 (5)

$$\dot{z}_t = u_t(z \mid \epsilon) = \frac{\dot{a}_t}{a_t} z - \frac{\dot{a}_t b_t - a_t \dot{b}_t}{a_t} \epsilon \tag{9}$$

现在，考虑*信噪比* $\lambda_t := \log \frac{a_t^2}{b_t^2}$。随着 $\lambda_t' = 2\left(\frac{\dot{a}_t}{a_t} - \frac{\dot{b}_t}{b_t}\right)$，我们可以将方程 (9) 改写为

$$u_t(z \mid \epsilon) = \frac{\dot{a}_t}{a_t} z - \frac{1}{2} \lambda_t' b_t \epsilon \tag{10}$$

接下来，我们使用公式 (10) 将公式 (8) 重新参数化为噪声预测目标：

$$L_{CFM} = \mathsf{E}_{t,p(z \mid \epsilon),p(\epsilon)}\left\| \frac{\dot{a}_t}{a_t} z + \frac{1}{2} \lambda_t' b_t \epsilon - v_{(\Theta)}(z, t) \right\|_2^2 \tag{11}$$

$$= \mathsf{E}_{t,p(z \mid \epsilon),p(\epsilon)}\left[-\frac{1}{2} \lambda_t' \right] \left\| \epsilon_{(\Theta)}(z, t) - \epsilon \right\|_2^2 \tag{12}$$

其中我们定义了 $\epsilon_\Theta := -\frac{2}{\lambda_t' b_t}(v_\Theta - \dot{a}_t x)$

我们可以推导出各种加权损失函数，它们提供了通向理想解的信号，但可能会影响优化轨迹。为了统一分析不同的方法，包括经典的扩散公式，我们可以将目标写成以下形式（仿效 Kingma & Gao (2023)）：

$$L_w(x_0) = -\frac{1}{2} \mathsf{E}_{t \sim U(t), \epsilon \sim N(0,I)}\left[w_t \lambda_t' \left\| \epsilon_\Theta(z_t, t) - \epsilon \right\|^2 \right] 、$$

其中 $w_t = -\frac{1}{2} \lambda_t' b_t^2$对应 L_{CFM}.

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，下面将对其进行简要说明。

整流 整流（RF）（Liu 等人，2022；Albergo & Vanden-Eijnden, 2022; Lipman et al.
数据分布和标准正态分布，即

$$z_t = (1 - t)x_0 + t\epsilon 、 \tag{13}$$

并使用$_{L(CF)(M)}$，对应于 $w_t^{\mathrm{RF}} = \frac{t}{1-t}$
网络输出直接参数化了速度 v_Θ。

EDM EDM（Karras 等人，2022 年）使用的前向过程形式为

$$z_t = x_0 + b_t \epsilon \tag{14}$$

其中 (Kingma & Gao, 2023) $b_t = \exp F^{-1}(t \mid P_{(m)}, P^{(2)})$
F^{-1}是 no rmal 分布- 与 $N(P_m, P_s^2)$ 均值为 P_m，方差为 P_s^2。请注意结果为

$$\lambda_t \sim N(-2P_m, (2P_s)^2) \quad \text{for } t \sim U(0, 1) \tag{15}$$

通过 F **预测对**网络进行参数化（Kingma 和 Gao，2023 年；Karras 等人，2022 年），损失为可写成 L_w^{EDM}，带

$$w_t^{\mathrm{EDM}} = N(\lambda_t \mid -2P_m, (2P_s)^2)(e^{-\lambda_t} + 0.5^2) \tag{16}$$

余弦 （Nichol & Dhariwal，2021 年）提出了一种前瞻性的过程的形式

$$z_t = \cos\frac{\pi}{2} t\, x_0 + \sin\frac{\pi}{2} t\, \epsilon . \tag{17}$$

结合ϵ参数化和损耗，这个对应于权重 $w_t = \mathrm{sech}(\lambda_t/2)$。当

请注意，在引入随时间变化的权重时，上述目标的最佳值不会发生变化。因此

结合 **v 预测**损失（Kingma & Gao，2023 年），权重为 $w_t = e^{-\lambda(t)/2}$。

(LDM-)Linear LDM（Rombach 等人，2022 年）使用 DDPM 计划表（Ho 等人，2020 年）的修正版。二者都是方差保持计划，即 $b_t = 1 - a^2$，并对离散时间步 $t = 0, \ldots, T-1$ 的 $a_{(t)}$ 进行去精细化处理。Q_t 作为 $t = (\prod_{s=0}^{t} (1 - \beta))^{\frac{1}{2}}$ 的扩散系数 β

对于给定的边界值 β_0 和 β_{T-1}，DDPM 使用 $\beta_t = \beta_0 + (t)(\beta_{T-1} - \beta_{(0)})$ 而 LDM 使用 $\beta_{(t)} = \sqrt{(\beta)} + t(\sqrt{\beta_{T-1}} - \sqrt{\beta_0})^2$

3.1. 射频模型的定制信噪比采样器

射频损耗 v_Θ 在所有时间步上均匀地在 [0, 1] 中。然而，直观地说，当 t 在 $[0]$ 和 $[1]$ 之间时，速度预测目标 $\epsilon - x_0$ 的难度更大。

[0，1]，因为对于 $t = 0$，最优预测是 p_1 的均值，而对于 $t = 1$，最优预测是 $p_{(0)}$ 的均值。一般来说，将 t 的分布从常用的均匀分布 $U(t)$ 改为分布

密度 $\pi(t)$ 等于加权损失 L_{w^π}，其中

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

因此，我们的目标是通过更频繁地对中间时间步进行采样，来提高中间时间步的权重。接下来，我们将描述用于训练模型的时间步密度 $\pi(t)$。

对数正态采样 对数正态分布（Atchison & Shen，1980 年）是一种更重视中间步骤的分布。其密度为

$$\pi_{\ln}(t; m, s) = \frac{1}{(s)\sqrt{(2)}\,\overline{(\pi)}(t)\,(1-(t))} \exp\frac{(\text{logit}(t) - m)^{(2)}}{2s2}, \tag{19}$$

其中 $\text{logit}(t) = \log t$，有一个位置参数 m，以及尺度参数 s。位置参数使我们能够使训练时间步偏向数据 $p_{(0)}$（负 m）或噪声 $p_{(1)}$（正 m）。如图 8 所示，比例参数控制着分布的宽窄。

在实践中，我们从一个正态分布 $u(u; m, s)$ 中采样，并通过标准 logistic 函数进行映射。

重尾模式采样 logit-normal 分布在端点 0 和 1 处总是消失的。为了研究这是否会对性能产生不利影响，我们还使用了一个时间步采样分布，其密度严格正对 [0，1]。对于规模参数 s，我们定义

$$f_{\text{mode}}(u; s) = 1 - u - s \cdot \cos^{(2)}\frac{\pi}{2}u - 1 + u \tag{20}$$

对于 $1 - \leq s \leq 2$，该函数是单调的，我们可以从隐含密度 $\pi_{(\text{模式})}(t; s)$ 中采样 $= f^{-1}_{\text{模式}}(t)$。如图 8 所示，比例参数控制取样过程中对中点（正 s）或端点（负 s）的偏好程度。该公式还包括一个**均匀加权** $\pi_{\text{mode}}(t; s = 0) = U(t)$ for $s = 0$，这在以前的**整流研究中**得到了广泛应用（Liu 等人，2006；Liu 等人，2007；Liu 等人，2008；Liu 等人，2009）。在以前的整流理论研究中被广泛使用（Liu et al、2022 年；Ma 等人，2024 年）。

CosMap 最后，我们还在射频环境中考虑了第 3 节中的*余弦*时间表（Nichol & Dhariwal, 2021 年）。

具体而言，我们正在寻找一个映射 $f : u \to f(u) = t, u \in [0, 1]$，从而使对数 snr 与余弦的对数 snr 匹配附表：$2\log(\frac{\cos(u)}{\sin(\frac{\pi}{2}u)}) = 2\log(\frac{1-f(u)}{f(u)})$。求出 f，得 $\sim U(u)$

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

得出密度

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

4. 文本到图像架构

对于文本条件下的图像采样，我们的模型必须同时考虑文本和图像两种模式。我们使用预训练模型来推导合适的表征，然后描述扩散骨干网的架构。图 2 是这一架构的概览。

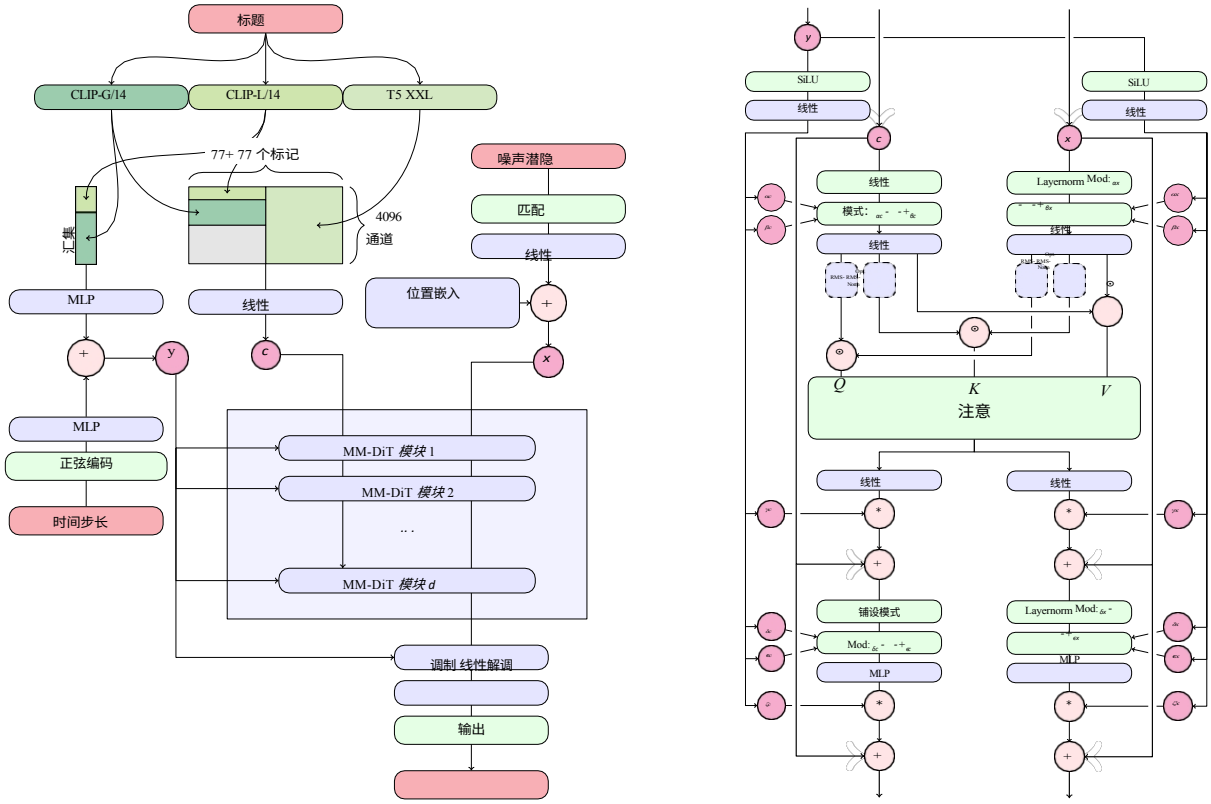
我们的一般设置遵循 LDM（Rombach 等人，2022 年）的潜空间中训练文本到图像模型。

预训练的自动编码器。与将图像编码为潜在表征类似，我们也沿用了以前的方法（Saharia 等人，2022b；Balaji 等人，2022），并对文本进行编码。位置参数使我们能够使用预训练的冻结文本模型来调节 c。详细信息可参见附录 B.2。

多模态扩散骨干网 我们的架构以 DiT（Peebles & Xie，2023 年）架构为基础。DiT 只考虑类别条件下的图像生成，并使用一种调制机制，将扩散过程的时间步和类别标签作为网络的条件。同样，我们将时间步 t 和 c_{vec} 的嵌入作为调制机制的输入。然而，由于集合文本表示法只保留了文本输入的粗粒度信息（Podell 等人，2023 年），因此网络还需要序列*表示法* $_{\text{cctxt}}$ *的*信息。

我们构建了一个由文本和图像输入嵌入组成的序列。具体来说，我们添加位置嵌入编码，并平铺 2× 2 个潜在像素代表补丁。对 $x \in \mathbb{R}^{h \times w \times c}$ 的补丁编码序列进行编码。

长度 $_2h\,_2^1\,ww$。在将该补丁编码和文本编码 c_{ctxt} 嵌入到一个共同维度后，我们



(a) 所有组件概览。

(b) 一个 *MM-DiT* 模块

*图 2.***我们的模型架构。**连接用⊙ 表示，元素相乘用∗ 表示。可以添加 Q 和 K 的 RMS-Norm 来稳定训练运行。放大查看效果更佳。

连接两个序列。然后，我们按照 DiT 的方法，应用调制注意力和 MLP 序列。

由于文字嵌入和图像嵌入在概念上有很大不同，因此我们为两种模式分别设置了两套权重。如图 2b 所示，这相当于为每种模态设置了两个独立的变换器，但在注意力操作时将两种模态的序列连接在一起，这样两种表征都可以在各自的空间中工作，同时也将另一种空间考虑在内。

在缩放实验中，我们以模型的深度 d（即注意力区块的数量）作为模型大小的参数，设置隐藏大小为 $64\,d$（在 MLP 区块中扩展为 $4 \cdot 64\,d$ 通道），注意力头的数量等于 d。

5. 实验

5.1. 改进整流

我们的目标是了解等式 1 中的归一化流量无模拟训练方法中哪种最有效。为了对不同方法进行比较，我们对优化算法、模型结构、数据集和采样器进行了控制。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，我们需要能对不同方法进行比较的评估指标。我们在 Ima- geNet（Russakovsky 等人，2014 年）和 CC12M（Changpinyo 等人，2021 年）上训练模型，并在训练过程中使用验证损失、CLIP 分数（Radford 等人，2021 年；Hessel 等人，2021 年）和 FID（Heusel 等人，2017 年）在不同的采样器设置（不同的指导尺度和采样步骤）下评估模型的训练和 EMA 权重。我们根据（Sauer 等人，2021 年）提出的 CLIP 特征计算 FID。所有指标都在 COCO-2014 valida- tion split（Lin 等人，2014 年）上进行了评估。有关训练和采样超参数的全部细节见附录 B.3。

5.1.1. 结果

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

- 采用线性（`eps/linear`，`v/linear`）和余弦（`eps/cos`，`v/cos`）计划的ϵ和 $\mathbf{v}$ **预测**损失。
- 射频损耗与 $\pi_{(\text{mode})}(t; s)$（`rf/mode(s)`），其中 s 的 7 个值在$-1$ 和 1.75 之间均匀选择，并在 1 和 1.75 之间均匀排列。

	秩的平均值		
变量	所有	5 步	50 步
rf/lognorm(0.00, 1.00)	1.54	1.25	1.50
RF/LOGNORM(1.00, 0.60)	2.08	3.50	2.00
RF/LOGNORM(0.50, 0.60)	2.71	8.50	1.00
转速/模式（1.29）	2.75	3.25	3.00
RF/LOGNORM(0.50, 1.00)	2.83	1.50	2.50
ε/线性	2.88	4.25	2.75
红外线/模式(1.75)	3.33	2.75	2.75
转速/宇宙图	4.13	3.75	4.00
教育模式（0.00，0.60）	5.63	13.25	3.25
rf	5.67	6.50	5.75
v/ 线性	6.83	5.75	7.75
EDM(0.60, 1.20)	9.00	13.00	9.00
v/cos	9.17	12.25	8.75
教育管理/计算机	11.04	14.25	11.25
教育管理/广播	13.04	15.25	13.25
教育模式（-1.20，1.20）	15.58	20.25	15.00

*表 1.***变体的总体排名。**在排序时，我们对 EMA 权重和非 EMA 权重、两个数据集和不同的抽样设置进行了非优势排序。

	图像网络		CC12M	
变体	CLIP	FID	CLIP	FID
rf	0.247	49.70	0.217	94.90
EDM(-1.20, 1.20)	0.236	63.12	0.200	116.60
ε/线性	0.245	48.42	0.222	*90.34*
v/cos	0.244	50.74	0.209	97.87
v/linear	0.246	51.68	0.217	100.76
rf/lognorm(0.50, 0.60)	**0.256**	80.41	0.233	120.84
rf/mode(1.75)	*0.253*	**44.39**	0.218	94.06
RF/LOGNORM(1.00, 0.60)	0.254	114.26	**0.234**	147.69
rf/lognorm(-0.50, 1.00)	0.248	45.64	0.219	**89.70**
rf/lognorm(0.00, 1.00)	0.250	*45.78*	*0.224*	89.91

*表 2.***不同变体的指标。**25 个采样步骤下不同变体的 FID 和 CLIP 分数。我们突出显示了**最佳**、次佳和*三佳*条目。

另外，$s = 1.0$ 和 $s = 0$ 对应统一时间步采样（`rf/mode`）。

- 射频损耗 $\pi_{(\ln)}(t; m, s)$（`rf/lognorm(m, s)`），网格中（m, s）有 30 个值，m 在 −1 和 1 之间均匀分布，s 在 0.2 和 2.2 之间均匀分布。
- 射频损耗 $\pi_{(\text{CosMap})}(t)$（`rf/cosmap`）。
- EDM（`edm`(P_m, P_s)），P_m的 15 个值在 1.2 和 1.2 之间均匀选择，$P_{(s)}$的 15 个值在 0.6 和 1.8 之间均匀选择。请注意，$P_m, P_s = (1.2, 1.2)$ 相当于 (Karras et al., 2022) 中的参数。
- 放电解调，其时间表与 `rf` 的对数-SNR 加权（`edm/rf`）相匹配，与 `v/cos` 的对数-SNR 加权（`edm/cos`）相匹配。

对于每次运行，我们都会选择使用 EMA 权重评估时验证损失最小的步骤，然后收集在 6 种不同采样器设置下获得的 CLIP 分数和 FID。

在有 EMA 权重和无 EMA 权重的 24 个采样器组合中

对于采样器设置、EMA 权重和数据集选择的所有 24 种组合，我们使用非优势排序算法对不同的公式进行排序。为此，我们根据 CLIP 和 FID 分数反复计算帕累托最优变体，为这些变体分配当前迭代指数，移除这些变体，然后继续计算其余变体，直到所有变体都排上名次。最后，我们对 24 个不同的控制设置进行平均排名。

结果见表 1。表 1 中，我们只显示了使用不同超参数评估的变体中表现最好的两个变体。我们还显示了对 5 步和 50 步采样器设置进行平均后的排名。

我们发现 `rf/lognorm(0.00,1.00)` 始终保持着良好的排名。它优于统一时间步采样的整流公式 (`rf`)，从而证实了我们的假设，即中间时间步更重要。

更重要。在所有变体中，*只有*整流式

与之前使用的 LDM-Linear（Rombach 等人，2022 年）公式（`eps/linear`）相比，使用修改时间步采样的公式表现更好。

我们还观察到，有些变体在某些情况下表现良好，但在其他情况下则较差，*例如*，`rf/lognorm(0.50, 0.60)` 在 50 个采样步骤时表现最佳，但在 5 个采样步骤时则差得多（平均排名 8.5）。我们在表 2 中观察到这两个指标的类似表现。2.第一组显示了在 25 个采样步长的两个数据集上的代表性变体及其指标。下一组显示了获得最佳 CLIP 和 FID 分数的变体。除了 `rf/mode(1.75)`，这些变体通常在一个指标上表现非常好，但在另一个指标上相对较差。与此相反，我们再次观察到 `rf/lognorm(0.00,1.00)` 在不同指标和数据集上都取得了良好的表现，它在四次得分中两次获得第三名，一次获得第二名。

最后，我们在图 3 中说明了不同公式的定性行为，我们用不同的颜色表示不同的公式组（`edm`、`rf`、`eps` 和 `v`）。整流公式一般表现良好，与其他公式相比，在减少采样步骤数时，其性能下降较少。

5.2. 改进特定模态表示

在上一节中，我们找到了一种公式，它不仅能让整流模型与线性扩散公式（LDM-Linear）（Rom- bach 等人，2022 年）或 EDM（Karras 等人，2022 年）等成熟的扩散公式竞争，甚至还能超越它们。通过

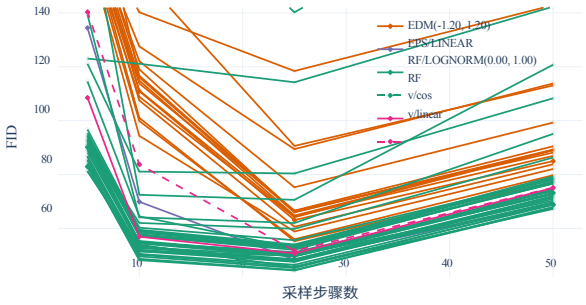


图 3. **整流取样效率高。**当采样步数较少时，整流公式的性能优于其他公式。对于 25 步及更多步骤，只有 `rf/lognorm(0.00,1.00)` 比 `eps/linear` 更有竞争力。

指标	4 步	8 步	16 步
FID ( )↓	2.41	1.56	**1.06**
知觉相似性 ( )↓	0.85	0.68	**0.45**
SSIM ( )↑	0.75	0.79	**0.86**
PSNR (↑)	25.12	26.40	**28.62**

表 3. **改进的自动编码器。**不同信道配置下的重建性能指标。所有模型的下采样系数均为 $f = 8$。

因此，我们算法的最终性能不仅取决于训练配方，还取决于通过神经网络进行的参数化以及我们使用的图像和文本表示的质量。在接下来的章节中，我们将介绍如何改进所有这些部分，然后在第 5.3 节中对我们的最终方法进行缩放。

5.2.1. 改进的自动编码器

潜像扩散模型通过在预训练自动编码器（Rombach 等人，2022 年）的潜像空间中运行实现高效率，该自动编码器将输入 RGB $x \in \mathbb{R}^{H \times W \times 3}$ 映射到低维空间 $z = E(x) \in \mathbb{R}^{h \times w \times d}$。这种自动编码器的重建质量为潜在扩散训练后可实现的图像质量提供了上限。与 Dai 等人（2023 年）的研究类似，我们发现增加潜通道的数量 d 能显著提高重建性能，见表 3。直观地说，预测更多 d 的潜变量是一项更困难的任务，因此容量更大的模型应该能够在更多 d 的情况下表现更好，最终实现更高的图像质量。我们在图 7 中证实了这一假设，在图 7 中，我们看到 $d = 16$ 自动编码器在样本 FID 方面表现出更好的扩展性能。因此，在本文的其余部分，我们选择 $d = 16$。

5.2.2. 改进字幕

Betker 等人（2023 年）证明，合成生成的字幕可以极大地改进大规模训练的文本到图像模型。这是由于合成的标题通常比较简单。

	原始字幕	50/50 混合
	成功率 [%]	成功率 [%]
色彩归属	11.75	24.75
颜色	71.54	68.09
位置	6.50	18.00
计数	33.44	41.56
单个物体	95.00	93.75
两个物体	41.41	52.53
总分	43.27	49.78

表 4. **改进的字幕。**使用 50/50 的合成（通过 CogVLM（Wang 等人，2023 年））和原始字幕混合比例可以提高文本到图像的性能。通过 GenEval（Ghosh 等人，2023 年）基准进行评估。

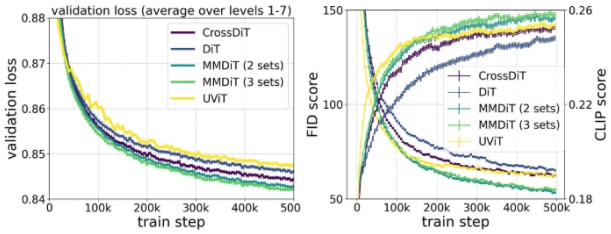


图 4：模型架构的训练动态 **模型架构的训练动态。***DiT*、*CrossDiT*、*UViT* 和 *MM-DiT* 在 CC12M 上的比较分析，重点是验证损失、CLIP 分数和 FID。我们提出的 *MM-DiT* 在所有指标上都表现出色。

大规模图像数据集所附带的人工生成的注释过于关注图像主体，通常忽略了描述背景或场景构成的细节，或者，如果适用的话，忽略了显示的文本（Betker 等人，2023 年）。我们沿用他们的方法，使用现成的、最先进的视觉语言模型 *CogVLM*（Wang 等人，2023 年）为我们的大规模图像数据集创建合成注释。由于合成标题可能会导致文本到图像模型遗忘 VLM 知识语料库中不存在的某些概念，因此我们使用了 50% 原始标题和 50% 合成标题的比例。

为了评估在这种字幕混合情况下进行训练的效果，我们对两个 $d = 15$ *MM-DiT* 模型进行了 25 万步的训练，其中一个只对原始字幕进行训练，另一个对 50/50 混合字幕进行训练。我们使用 GenEval 基准（Ghosh 等人，2023 年）对训练后的模型进行了评估，见表 4。结果表明，使用合成字幕训练的模型明显优于只使用原始字幕的模型。因此，我们在接下来的工作中使用合成/原始字幕各占一半的混合模式。

5.2.3. 改进的文本到图像骨架

在本节中，我们将现有的基于变压器的扩散骨干网与第 4 节中介绍的新型多模态变压器扩散骨干网 *MM-DiT* 进行性能比较。*MM-DiT* 专门设计用于

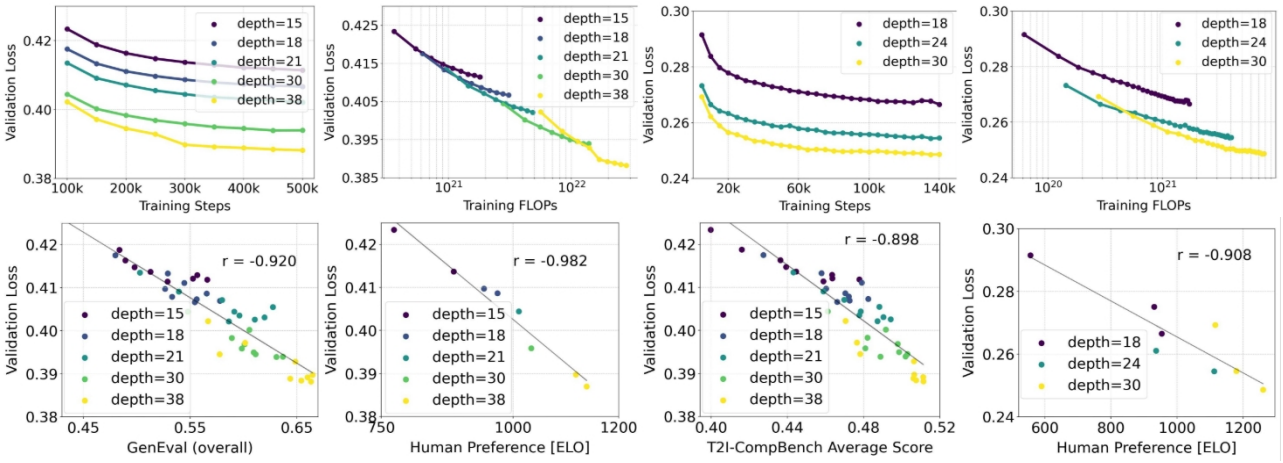


图 5. **缩放的定量影响。**我们分析了模型大小对性能的影响，整个过程中保持了一致的训练超参数。深度=38是个例外，为了防止发散，有必要在3×10^5步时调整学习率。(上图）图像模型（第 1 列和第 2 列）和视频模型（第 3 列和第 4 列）的验证损失随模型大小和训练步数的变化而平滑减小。(下图）验证损失是*模型整体性能的有力预测指标*。验证损失与整体图像评估指标之间存在明显的相关性，包括第 1 列 GenEval（Ghosh 等人，2023 年）、第 2 列人类偏好和第 3 列 T2I-CompBench（Huang 等人，2023 年）。对于视频模型，我们观察到验证损失与人类偏好（第 4 栏）之间存在类似的相关性。

我们将使用（两组）不同的可训练模型权重来处理不同的领域（这里是文本和图像标记）。更具体地说，我们按照第 5.1 节中的实验设置，比较了 DiT、CrossDiT（DiT，但对文本标记进行了交叉处理，而不是序列连接（陈等人，2023 年））和我们的 *MM-DiT* 在 CC12M 上的文本到图像性能。对于 *MM-DiT*，我们比较了有两套权重和三套权重的模型，其中后三套分别处理 CLIP（Radford 等，2021 年）和 T5（Raffel 等，2019 年）标记（*参见*第 4 节）。请注意，DiT（与第 4 节中的文本和图像标记连接）可以解释为 *MM-DiT* 的一种特例，所有模态都有一套共享权重。最后，我们将 UViT（Hoogeboom 等人，2023 年）架构视为广泛使用的 UNets 和变换器变体之间的混合体。

我们在图 4 中分析了这些架构的收敛行为：Vanilla DiT 的性能低于 UViT。交叉注意力 DiT 变种 CrossDiT 的性能优于 UViT，尽管 UViT 最初的学习速度似乎更快。我们的 *MM-DiT* 变体明显优于交叉注意和 vanilla 变体。我们观察到，使用三个参数集而不是两个参数集只会带来微小的收益（代价是参数数量和 VRAM 占用率的增加），因此在本工作的其余部分，我们选择了前者。

5.3. 规模化训练

接下来，我们想了解我们的模型在扩展时的行为。附录 C 介绍了缩放研究的前期准备工作，其中我们描述了在缩放训练数据（附录 C.1）和图像分辨率（附录 C.2）时确保高效稳定训练的必要步骤。附录 C.2）。然后，所有之前对扩散公式、架构和数据的考虑在最后一节达到了高潮，我们将模型扩展到 8B 参数。

5.3.1. 结果

在图 5 中，我们考察了按比例训练 *MM-DiT 的*效果。对于图像，我们使用预编码数据（参见附录 *C.1*）在 256^2像素分辨率下进行了大规模缩放研究，并训练了 500k 步不同参数数的模型。附录 C.1，批量大小为 4096。我们在 2 2 个斑块（Peebles & Xie，2023 年）上进行训练，每 50k 步在 CoCo 数据集（Lin 等人，2014 年）上报告验证损失。特别是，为了减少验证损失信号中的噪声，我们在 t（0，1）内等距离采样损失水平，并分别计算每个水平的验证损失。然后，我们对除最后一级（t= 1）以外的所有级别的损失进行平均。

同样，我们也在视频上对 *MM-DiT* 进行了初步的扩展研究。为此，我们从预训练的图像权重开始，并额外使用了 2 倍时间修补。我们效仿 Blattmann 等人（2023b）的做法，通过将时间轴折叠到批次轴来向预训练模型输入数据。在每个注意力层中，我们会重新排列视觉流中的表征，并在最后的前馈层之前，在空间注意力操作之后对所有时空标记添加完全注意力。我们的视频模型是在包含 16 帧 256^2 像素的视频上，以 512 的批量大小，经过 140k 步的训练而得到的。我们在 Kinetics 数据集（Carreira & Zisserman，2018 年）上报告了每 5k 步的验证损失。请注意，我们在图 5 中报告的视频训练 FLOP 值仅为视频训练的 FLOP 值，不包括图像预训练的 FLOP 值。

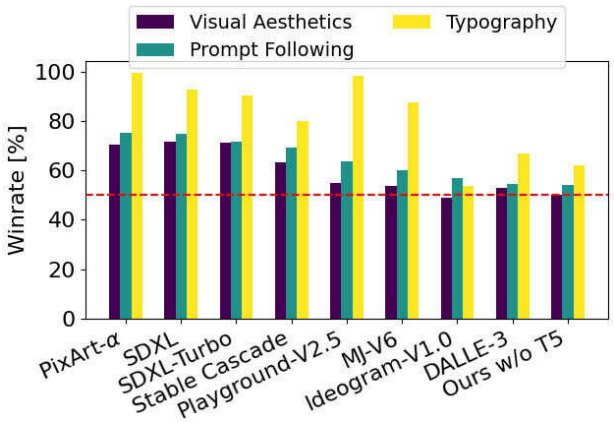


图 6. **针对当前封闭式和开放式 SOTA 生成图像模型的人类偏好评估。**与目前最先进的文本到图像模型相比，我们的 8B 模型在*视觉质量*、*提示跟踪*和*排版生成*等类别上的表现更胜一筹（Yu 等人，2022 年）。

在图像和视频领域，我们观察到随着模型规模和训练步骤的增加，验证损失也在平稳下降。我们发现验证损失与综合评估指标（Comp- Bench（Huang 等人，2023 年）、GenEval（Ghosh 等人，2023 年））和人类偏好高度相关。这些结果支持将验证损失作为衡量模型性能的一个简单而通用的指标。无论是图像模型还是视频模型，我们的结果都没有显示出饱和状态。

图 14 说明了长时间训练大型模型对样本质量的影响。表 5 显示了 GenEval 的全部结果。表 5 显示了 GenEval 的全部结果。当应用附录 C.2 中介绍的方法并提高训练图像的分辨率时，我们的最大模型在大多数类别中都表现出色，总分超过了 DALLE 3（Betker 等人，2023 年），这是目前最先进的提示理解模型。

在 Parti-prompts 基准（Yu 等人，2022 年）的*视觉美学*、*提示遵循*和*排版*生成类别的人类偏好评估中，我们的 d= 38 模型优于当前的专有（Betker 等人，2023 年；ide，2024 年）和开放（Sauer 等人，2023 年；pla，2024 年；Chen 等人，2023 年；Pernias 等人，2023 年）SOTA 生成式图像模型，*参见图 6*。参见图 6。为了评估人类在这些类别中的偏好，我们向评分者展示了两个模型的成对输出结果，并要求他们回答以下问题：**提示如下***哪张图片看起来更能*代表*上面显示的文字并*忠实于*文字?*

视觉美学：*根据提示，哪张图片*质量更高、更美观*？* **排版：***哪张图片能更准确地显示/展示上述描述中指定的文字？拼写更准确者优先！忽略其他方面。*

模型	整体	对象 单一	两个	计数	颜色	位置	颜色 归属
minDALL-E	0.23	0.73	0.11	0.12	0.37	0.02	0.01
SD v1.5	0.43	0.97	0.38	0.35	0.76	0.04	0.06
PixArt-α	0.48	0.98	0.50	0.44	0.80	0.08	0.07
SD v2.1	0.50	0.98	0.51	0.44	0.85	0.07	0.17
DALL-E 2	0.52	0.94	0.66	0.49	0.77	0.10	0.19
SDXL	0.55	0.98	0.74	0.39	0.85	0.15	0.23
SDXL 涡轮增压发动机	0.55	**1.00**	0.72	0.49	0.80	0.10	0.18
IF-XL	0.61	0.97	0.74	*0.66*	0.81	0.13	0.35
DALL-E 3	0.67	0.96	0.87	0.47	*0.83*	**0.43**	*0.45*
我们的（深度=18），512^2	0.58	0.97	0.72	0.52	0.78	0.16	0.34
我们的（深度=24），512^{2a}	0.62	0.98	0.74	0.63	0.67	*0.34*	0.36
我们的（深度=30），512^{2a}	0.64	0.96	0.80	0.65	0.73	0.33	0.37
我们的（深度=38），512^2	*0.68*	0.98	0.84	*0.66*	0.74	0.40	0.43
我们的（深度=38），512^2/DPO	0.71	0.98	0.89	**0.73**	*0.83*	*0.34*	0.47
我们的（深度=38），1024^2/DPO	**0.74**	0.99	**0.94**	0.72	**0.89**	0.33	**0.60**

表 5. **GenEval 比较。**我们最大的模型（深度=38）在 GenEval（Ghosh et al.我们重点介绍了**最佳**、次佳和*三佳*项目。关于 DPO，请参见附录 C.3。

	相对 CLIP 分数降幅 [%]			
	5/50 级	10/50 步	20/50 步路径长度	
深度=15	4.30	0.86	0.21	191.13
深度=30	3.59	0.70	0.24	187.96
深度=38	2.71	0.14	0.08	185.96

表 6. **模型大小对采样效率的影响。**该表显示了在固定种子条件下，使用 50 个采样步骤评估 CLIP 分数时，相对于 CLIP 分数的性能下降情况。较大的模型可以使用较少的步骤进行采样，我们将此归因于鲁棒性的提高和更符合整流模型的直线路径目标，从而缩短了路径长度。路径长度的计算方法是对 50 步的$\left\|v_{(\theta)} - dt\right\|$求和。

最后，表 6 显示了一个耐人寻味的结果：更大的模型不仅性能更好，而且达到峰值所需的步骤也更少。

6. 结论

在这项工作中，我们对用于文本到图像合成的整流模型进行了缩放分析。我们为整流训练提出了一种新颖的时间步采样方法，这种方法比以往的潜扩散模型扩散训练公式更有优势，并在几步采样机制中保留了整流的有利特性。我们还展示了基于变压器的 *MM-DiT* 架构的优势，该架构考虑到了文本到图像任务的多模式特性。最后，我们对这一组合进行了扩展研究，将模型规模扩大到 8B 参数和 5 10^{22}个训练 FLOP。我们发现，验证损失的改进与现有的文本到图像基准以及人类偏好评估都有关联。这与我们在生成建模和可扩展多模态架构方面的改进相结合，实现了与最先进的专有模型相媲美的性能。扩展趋势没有显示出饱和的迹象，这让我们对未来继续提高模型性能充满信心。