One Transformer Fits All Distributions in Multi-Modal Diffusion at Scale

在多模态扩散中一个适合所有分布的转换器

摘要

本文提出了一个统一的**扩散框架**(称为*UniDiffuser*)，**它能在一个模型中拟合与一组多模态数据相关的所有分布。**而我们的关键思路是——边缘分布、条件分布和联合分布的学习扩散模型可以统一为预测扰动数据中的噪声。其中，扰动水平（即时间步长）对于不同的模态可以是不同的。

受到**统一视图**的启发，UniDiffuser 通过对原始扩散模型的最小修改，扰动所有模态的数据，而不是单一模态，**能同时学习到所有的分布**；输入不同模态的单个时间步长，**并预测所有模态的噪声**，而不是单一模态。总之，UniDiffuser 通过扩散模型的转换器参数化，以处理不同模态的输入类型的。

UniDiffuser 能在大规模成对的图像—文本数据上进行实现，能够通过设置适当的时间步长来执行图像、文本、文本到图像、图像到文本和图像—文本对的生成，**且没有额外的开销**。特别是，UniDiffuser 能够在所有任务中产生**在感知上真实的样本**，而其定量结果(例如 FID 和 CLIP 得分)不仅优于现有的通用模型，而且在代表性任务(例如，文本到图像生成任务)中与定制模型(例如，Stable Diffusion 和 DALL·E 2)相当。

我们的代码可在 https://github.com/thu-ml/unidiffuser 获得。（P1）

1. 介绍

最近，我们见证了一场由多模态数据生成建模的快速发展所推动的内容创作革命。特别是**扩散模型**(SohlDickstein et al., 2015; Ho et al., 2020; Song et al., 2021c)已经展示了创作高保真和多样化数据的、令人难以置信的能力(Ramesh et al., 2022; Saharia et al., 2022; Rombach et al., 2022; Ho et al., 2022a; Popov et al., 2021)，**其内容能与输入的文本条件很好地吻合。**

然而，这些生成模型只是被设计为**定制系统**，只允许执行单一的任务。但实际上，人类应该可以同时生成各种多模态内容，并使用任意的条件类型。例如，

艺术家可以根据文字、场景或想象力创作绘画，并利用语言能力生成照片的标题。总之，对于一个基于多模态数据的通用生成系统，**一个能够覆盖所有类型的多模态生成任务的统一训练框架**（见图 1）肯定是它的基本组件之一。（P1）

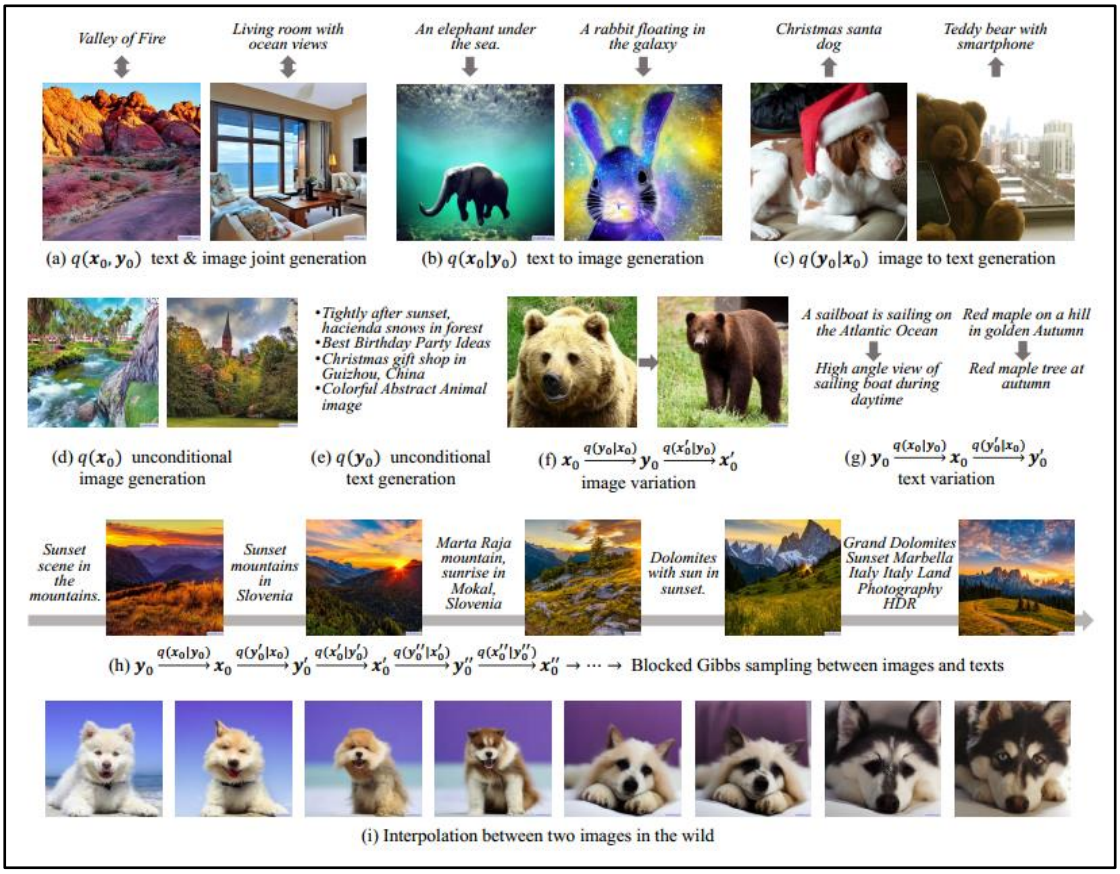

图 1：UniDiffuser 是通过用一个转换器(transformer)来拟合所有分布并处理各种任务的。在例子(a-e)中，UniDiffuser 可以直接进行联合生成、条件生成和无条件生成。在例子(f-g)中，图像变化和文本变化是利用 UniDiffuser 建模的两个条件分布的直接应用的。而例子(h)说明， UniDiffuser 可以执行阻塞吉布斯（blocked Gibbs）采样，以查看图像和文本是如何相互转换的**（能够实现在图文两个模态之间的来回跳跃）**。例子(i)则表明 UniDiffuser 能对真实的两张图像进行插值。（P2）

我们是从概率建模的角度拟合相应的分布来解决这一问题的。例如，从文本到图像的生成可以表述为学习**条件分布**$p(Image|Text)$。而拟合所有相关分布的**经典方法是隐式的**——它首先学习联合分布，然后通过额外的程序推断边缘和条件分布（例如，马尔科夫链蒙特卡罗(Srivastava & Salakhutdinov, 2012)）。但这在大规模多模态数据上是无法承受的(Schuhmann et al., 2022)。（P1）

相比之下，**本文提出了一个基于扩散的框架(称为 UniDiffuser)，它明确地在一个模型中适合所有相关的分布，而不引入额外的训练或推断开销。**我们的关键创新是：任何分布的学习扩散模型都可以统一为**预测扰动数据中的噪声**，其中扰动水平(即时间步长)对于不同的模态可以是不同的。例如，零级表示给定相应的模态条件生成，最大值表示忽略相应的模态无条件生成其他模态。受统一视图的启发，UniDiffuser 通过对原始扩散模型的最小修改，同时学习所有分布(Ho et al., 2020)（见图 2），即以所有模态而不是单一模态扰动数据，以不同模态输入单个时间步长，并预测所有模态而不是单一模态的噪声。当然，UniDiffuser 也能够以与指定扩散模型相同的方式执行各种生成（参见图 1）。此外，UniDiffuser 也可以免费执行无分类器指导(Ho & Salimans, 2021)，以提高条件生成和联合生成中的样本质量，因为 UniDiffuser 已经对边际分布进行了建模。（P1 & P2）

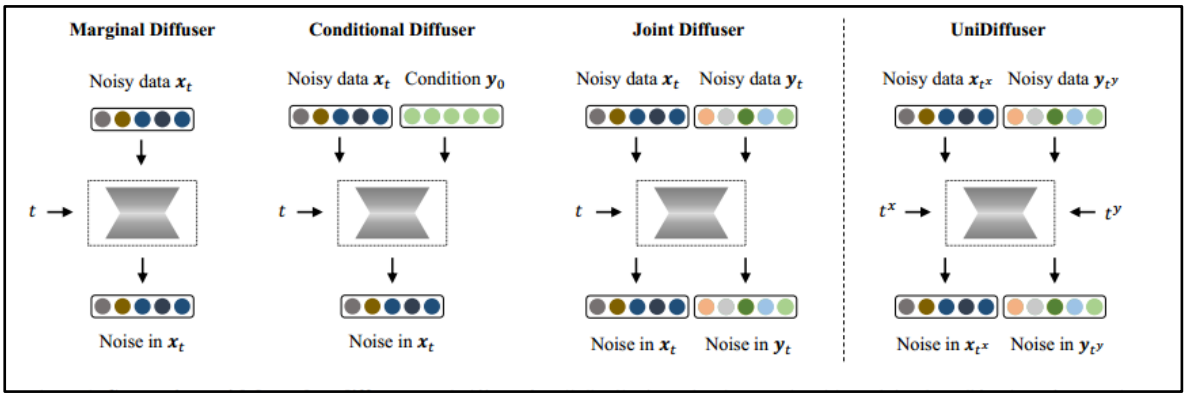

图 2：UniDiffuser 与一些定制转换器的比较。UniDiffuser 能通过对 Ho 等的最小修改同时拟合所有分布(2020)。特别是，当通过适当地设置时间步长(或噪声)时，它能退化到指定的扩散模型。（P2）

除了**概率建模框架**，**一个可以处理不同形式输入类型的统一架构**是通用生成系统的另一个基本组成部分。值得注意的是，**Transformer** 的出现(Vaswani et al., 2017; Dosovitskiy et al., 2021)及其在生成建模中的应用(Bao et al., 2022a)提供了一种很有前途的解决方案来捕获模式之间的相互作用。**当然，UniDiffuser 使用了一个基于 transformer 的主干。**（P2）

我们在潜在空间((Rombach et al., 2022)中实现了 UniDiffuser，并为图像提供了额外的 CLIP 编码器(Radford et al.,2021)，同时也为大规模图像—文本数据上

的文本提供了 GPT-2 (Radford et al., 2019)解码器(Schuhmann et al.,2022)。

总之，UniDiffuser 能够通过设置适当的时间步长来执行图像、文本、文本到图像、图像到文本和图像—文本对的生成，而没有额外的开销。特别是，UniDiffuser 能够在所有任务中产生感知上真实的样本，其定量结果(例如 FID 和 CLIP 评分)不仅优于现有的通用模型，而且在代表性任务(例如文本到图像生成)中与相应的定制模型(Stable Diffusion and DALL·E 2)相当。（P2）

2. 背景

扩散模型 (Sohl-Dickstein et al., 2015; Ho et al.,2020)是通过逐步向数据$x_0 \sim q(x_0)$注入噪声来扰动数据的，它由马尔可夫链形式化：

$$q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) = \prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}),$$
$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t|\sqrt{\alpha_t}\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I}),$$

where β_t is the noise schedule and $\alpha_t = 1 - \beta_t$.

而通过反转这一过程我们就可以生成数据，其中反向跃迁$q(x_{t-1}\,|x_t\,)$近似值为高斯模型$p(x_{t-1}\,|x_t\,) = \mathcal{N}(x_{t-1}|\mu_t(x_t), \sigma_t^2 I)$。而如 Bao 等人(2022c)所示，它在最大似然估计下的最优平均值为：

$$\boldsymbol{\mu}_t^*(\boldsymbol{x}_t) = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\mathbb{E}[\boldsymbol{\epsilon}^x|\boldsymbol{x}_t]\right), \quad (1)$$

其中，$\overline{\alpha_t} = \prod_{i=1}^{t} \alpha_i$ 和ϵ^x为注入到x_t的标准高斯噪声。而为了估计条件期望$\mathbb{E}[\epsilon^x|x_t]$，需要采用噪声预测网络$\epsilon_\theta(x_t, t)$来使回归损失最小，如下所示：

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{\epsilon}^x}\|\boldsymbol{\epsilon}^x - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)\|_2^2, \quad (2)$$

其中，t 是从{1,2，…， T}的均匀采样，$x_t = \sqrt{\overline{\alpha_t}}x_0 + \sqrt{1-\overline{\alpha_t}}\epsilon^x$。而根据$L2$回归损失的性质，最优噪声预测网络需满足$\epsilon_\theta(x_t, t) = \mathbb{E}[\epsilon^x|x_t]$。而又因为 Eq.(2)也等价于去噪分数匹配损失(Vincent, 2011)，因此最优噪声预测网络也满足$\epsilon_\theta(x_t, t) = -\sqrt{\beta_t}\nabla \log\, \mathrm{q}(x_t)$，其中$\mathrm{q}(x_t)$是扰动数据在时间步 t 的分布。

条件扩散模型。在条件生成的情况下，我们有配对数据$(x_0, y_0) \sim q(x_0, y_0)$，并且我们还想要建模条件数据分布$q(x_0, y_0)$以$y_0$为条件的逆向过程的高斯模型为$p(x_{t-1}|x_t, y_0) = \mathcal{N}(x_{t-1}|\mu_t(x_t, y_0), \sigma_t^2 I)$。与式(1)类似，它在最大似然估计下的最优均值为：

$$\boldsymbol{\mu}_t^*(\boldsymbol{x}_t, \boldsymbol{y}_0) = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\mathbb{E}[\boldsymbol{\epsilon}^x|\boldsymbol{x}_t, \boldsymbol{y}_0]\right). \quad (3)$$

为了估计$\mathbb{E}[\epsilon^x|x_t, y_0]$，还需要采用以$y_0$为条件的噪声预测网络，来使回归损失最小：

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{y}_0,\boldsymbol{\epsilon}^x} \|\boldsymbol{\epsilon}^x - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \boldsymbol{y}_0, t)\|_2^2.$$

此外，为了提高条件扩散模型的样本质量，(Ho & Salimans, 2021)提出了**无分类器引导(CFG)**。具体来说，它是通过线性组合条件模型和无条件模型进行采样的：

$$\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \boldsymbol{y}_0, t) = (1+s)\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \boldsymbol{y}_0, t) - s\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t), \quad (4)$$

其中s为指导标度。条件和无条件模型通过引入空令牌$\emptyset$来共享参数，例如$\epsilon_\theta(x_t, t) = \epsilon_\theta(x_t, y_0 = \emptyset, t)$。（P3）

3. 方法

在 3.1 节中，介绍了 UniDiffuser，这是一个单一的扩散模型，用于捕获由多模态数据同时确定的边缘分布、条件分布和联合分布。在 3.2 节中，演示了如何在 UniDiffuser 的条件采样和联合采样中不受约束的执行无分类器引导(CFG)。而为了简单起见，**本文主要讨论双模态数据**，但是 UniDiffuser 能够很容易地扩展到更多模态的数据中。（P3）

3.1. UniDiffuser：一个适合所有分布的扩散

在形式上，假设我们有两种从分布$q(x_0, y_0)$中采样的数据。我们的目标是设计一个基于扩散的模型，能够捕获所有由$q(x_0, y_0)$决定的相关分布，即边际分布$q(x_0)$和$q(y_0)$，条件分布$q(x_0|y_0)$和$q(y_0|x_0)$，联合分布$q(x_0, y_0)$。（P4）

我们注意到，用扩散模型学习一个分布等价于在噪声上估计一个条件期望。特别地，对边际分布$q(x_0)$建模相当于把估计注入到x_t噪声的条件期望中，即，根据式(1)，$\mathbb{E}[\epsilon^x|x_t]$。同样地，对条件分布$q(x_0|\,y_0)$和联合分布$q(x_0,y_0)$建模需要估计的关键量分别是$\mathbb{E}[\epsilon^x|x_t,y_0]$ （见式(3)）和$\mathbb{E}[\epsilon^x,\epsilon^y|x_t,y_t]$。（P4）

一个关键的发现是，以上所有的条件期望都可以统一为$\mathbb{E}[\epsilon^x,\epsilon^y|x_{t^x},y_{t^y}]$的一般形式，其中$t^x$和$t^y$是两个可以不同的时间步长，而$x_{t^x}$和$y_{t^y}$是相应的扰动数据。特别地，最大时间步$T$意味着将其边缘化。也就是说，设$t^y=T$，我们有$\mathbb{E}[\epsilon^x|x_t,y_T]\approx\mathbb{E}[\epsilon^x|x_{t^x}]$[1]，这对应于边际分布$q(x_0)$。

类似地，零时间步长意味着对相应的模态进行条件调节，而固定的时间步长意味着联合采样两种模态。因此，在形式上$\mathbb{E}[\epsilon^x|x_{t^x},y_0]$对应条件分布$q(x_0|\,y_0)$，通过设置$t^y=0$。而$\mathbb{E}[\epsilon^x,\epsilon^y|x_t,y_t]$对应于联合分布$q(x_0,y_0)$，通过设置$t^x=t^y=t$。此外，我们还可以描述$q(x_0|\,y_{t^y})$和$q(\,y_0|x_{t^x})$对所有$t^y$、$t^x$和生成数据条件的输入，通过估算$\mathbb{E}[\epsilon^x,\epsilon^y|x_{t^x},y_{t^y}]$。（P4）

受统一视图的启发，我们学习了所有$0\le t^x,t^y\le T$的$\mathbb{E}[\epsilon^x,\epsilon^y|x_{t^x},y_{t^y}]$，来模拟由$q(x_0,y_0)$决定的所有相关分布。具体而言，我们采用联合噪声预测网络$\epsilon_\theta(x_{t^x},y_{t^y},t^x,t^y)$，通过最小化以下回归损失函数来预测注入到$x_{t^x}$和$y_{t^y}$的噪声，类似于 Ho 等人(2020)的研究：

$$\mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{y}_0,\boldsymbol{\epsilon}^x,\boldsymbol{\epsilon}^y,t^x,t^y}\|\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t^x},\boldsymbol{y}_{t^y},t^x,t^y)-[\boldsymbol{\epsilon}^x,\boldsymbol{\epsilon}^y]\|_2^2,\quad(5)$$

其中(x_0,y_0)是一种随机的数据点，[,]表示连接，ϵ^x和ϵ^y表示从标准高斯分布的抽样，t^x,t^y是从独立$\{1,2,\ldots.,T\}$的均匀采样。**总之，我们称我们的方法为UniDiffuser，因为它以统一的方式捕获多个分布。**我们在附录 B 中给出了训练算法。（P4）

式(5)中的目标与式(2)中的原 DDPM 一样简单，而且对于一次参数的更新，UniDiffuser 只需要对多个任务(即分布)进行一次正向向后计算，其效率与原 DDPM 相同。由于两个独立的时间步长，UniDiffuser 的梯度估计的方差虽然略高于原始的 DDPM，但我们没有观察到 UniDiffuser 有较慢的收敛。（P4）

由于 UniDiffuser 试图通过一个联合噪声预测网络来拟合所有的分布，所以它要求骨干网可以处理模式之间的相互作用，并可用于大规模数据和多个任务。受 transformer 在大规模多模态表示学习上的出色表现的启发(Kim et al., 2021; Wang et al., 2022)，我们在 UniDiffuser 中使用基于 transformer 的神经网络，细节详见第 4.2 节。（P4）

给定一个单一的联合噪声预测网络，UniDiffuser 可以根据某个采样器进行无条件、条件和联合采样(采样算法见附录 B)。值得注意的是，通过适当地设置时间步长，UniDiffuser 的推理过程与定制模型相同。相比之下，学习单个联合分布(Srivastava & Salakhutdinov, 2012; Hu et al., 2022)在多模态数据上需要额外的程序(例如，马尔可夫链蒙特卡罗)来从边缘或条件分布中采样，这在**大规模多模态数据**上是无法承受的(Schuhmann et al., 2022)。（P4）

3.2. 不受约束的无分类器引导

无分类器引导(CFG) (Ho & Salimans, 2021)在采样过程中线性结合了条件模型和无条件模型（见式(4)）。该方法简单而有效地提高了扩散模型的样本质量和图像—文本对齐。但更值得一提的是，CFG 能直接适用于 UniDiffuser 的条件采样和联合采样且并不需要修改训练过程(结果见图 3)。（P5）

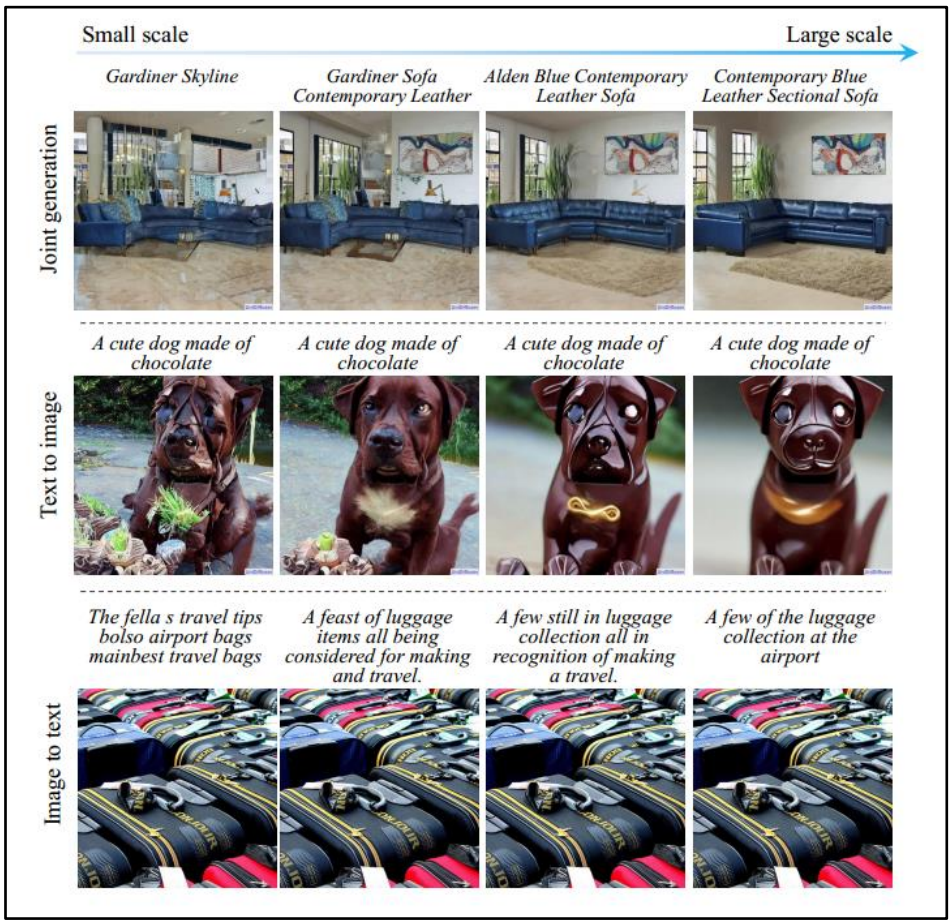

图 3：CFG 的影响。UniDiffuser 在联合采样和条件采样中无约束地使用了 CFG，提高了样本质量和图像—文本对齐，大尺度在 6 左右。

形式上，我们将ϵ_θ的输出表示为ϵ_θ^x和ϵ_θ^y，即$\epsilon_\theta = [\epsilon_\theta^x, \epsilon_\theta^y]$，而为了简单起见，我们省略了输入。此外，UniDiffuser 还可以在条件采样中无消耗执行 CFG，因为它能同时捕获条件和无条件模型。例如，我们可以生成条件为y_0的x_0，类似于式(4)：

$$\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}^{x}(\boldsymbol{x}_t, \boldsymbol{y}_0, t) = (1+s)\boldsymbol{\epsilon}_{\boldsymbol{\theta}}^{x}(\boldsymbol{x}_t, \boldsymbol{y}_0, t, 0) - s\boldsymbol{\epsilon}_{\boldsymbol{\theta}}^{x}(\boldsymbol{x}_t, \boldsymbol{\epsilon}^{y}, t, T),$$

其中，$\epsilon_\theta^x(x_t, y_0, t, 0)$和$\epsilon_\theta^x(x_t, \epsilon^y, t, T)$分别表示条件模型和无条件模型，而$s$为指导标度。且与原始 CFG 相比，UniDiffuser 不需要为参数共享指定空令牌。（P5）

CFG 也适用于联合采样。通过设置$t^x = t^y = t$，可以注意到联合评分模型可以等效地表示为条件模型的形式，如下所示：

$$\begin{aligned}\boldsymbol{\epsilon_\theta}(\boldsymbol{x}_t,\boldsymbol{y}_t,t,t)&\approx-\sqrt{\overline{\beta}_t}[\nabla_{\boldsymbol{x}_t}\log q(\boldsymbol{x}_t,\boldsymbol{y}_t),\nabla_{\boldsymbol{y}_t}\log q(\boldsymbol{x}_t,\boldsymbol{y}_t)]\\&=-\sqrt{\overline{\beta}_t}[\nabla_{\boldsymbol{x}_t}\log q(\boldsymbol{x}_t|\boldsymbol{y}_t),\nabla_{\boldsymbol{y}_t}\log q(\boldsymbol{y}_t|\boldsymbol{x}_t)],\end{aligned}$$

其中$q(x_t,y_t)$是同一噪声水平t下的扰动数据的联合分布。而受上述得分函数之间关系的启发，$\epsilon_\theta(x_t,y_t,t,t)$也可被视为近似于一对条件分数$\nabla_{x_t}\log\ \mathrm{q}(x_t|y_t)$和$\nabla_{y_t}\log\ \mathrm{q}(y_t|x_t)$。因此，我们可以通过将关节模型插值为相应的无条件模型来替换每个条件分数，如下所示：

$$\begin{aligned}&\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,\boldsymbol{y}_t,t)\\&=(1+s)\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,\boldsymbol{y}_t,t,t)-s[\boldsymbol{\epsilon}^x_{\boldsymbol{\theta}}(\boldsymbol{x}_t,\boldsymbol{\epsilon}^y,t,T),\boldsymbol{\epsilon}^y_{\boldsymbol{\theta}}(\boldsymbol{\epsilon}^x,\boldsymbol{y}_t,T,t)]\\&\approx-\sqrt{\overline{\beta}_t}[(1+s)\nabla_{\boldsymbol{x}_t}\log q(\boldsymbol{x}_t|\boldsymbol{y}_t)-s\nabla_{\boldsymbol{x}_t}\log q(\boldsymbol{x}_t),\\&\qquad\qquad(1+s)\nabla_{\boldsymbol{y}_t}\log q(\boldsymbol{y}_t|\boldsymbol{x}_t)-s\nabla_{\boldsymbol{y}_t}\log q(\boldsymbol{y}_t)],\end{aligned}$$

其中$\epsilon^x_\theta(x_t,\epsilon^y,t,T)$和$\epsilon^y_\theta(\epsilon^x,y_t,T,t)$代表无条件模型。此外，我们还在附录 C 中总结了 UniDiffuser 中 CFG 对于所有任务的公式。（P5）

4. UniDiffuser 的图像和文本

图像和文本是日常生活中最常见的两种形式。因此，验证 UniDiffuser 在这两种模式上的有效性具有一定的代表性。

我们的实现分为两个阶段(Rombach et al, 2022)（见图 4）。首先，我们通过图像和文本编码器将图像和文本转换为**连续的潜在嵌入x_0和y_0**，并引入两个解码器进行重建，如 4.1 节所述。其次，我们在潜在嵌入x_0和y_0上训练由 transformer 参数化的 UniDiffuser，如 4.2 节所述。（P5）

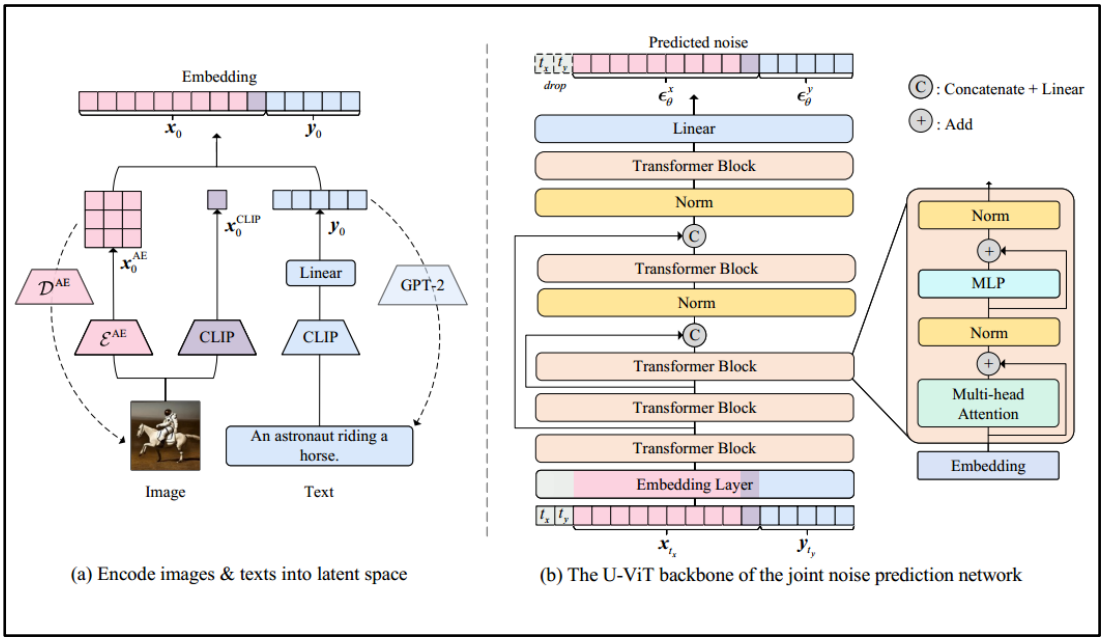

图 4：在图像—文本数据上实现 UniDiffuser。(a)图：首先，我们将图像和文本编码到潜在空间中。(b)图：其次，我们用 transformer(Bao et al, 2022a)参数化了 UniDiffuser，如图 2 所示，我们在潜在嵌入上进行模型训练。

4.1. 将图像和文本编码到潜在空间中

图像和文本编码器—解码器如图 4 (a)所示，下面我们提供了它们的详细信息。

图像编码器—解码器。图像编码器由两部分组成：第一部分是稳定扩散中使用的**图像自编码器**(Rombach et al, 2022)。我们使用它的编码器$\mathcal{E}^{AE}$来获得图像重建的嵌入x_0^{AE}。第二部分是图像 CLIP (Radford et al, 2021) (ViT-B/32)。它提取了一个 512 维的语义嵌入x_0^{CLIP}。所以，图像最终的潜在嵌入是这两部分输出的拼接，即$x_0 = [x_0^{AE}, x_0^{CLIP}]$。**根据经验，我们发现$x_0^{AE}$足以通过稳定扩散的图像解码器$D^{AE}$进行图像重建**，而额外的$x_0^{CLIP}$有助于理解图像到文本生成中的图像语义。此外，我们假设这两种嵌入的不同角色是由原始目标所固有的，即重建与文本的语义对齐。（P5）

文本编码器—解码器。对于文本编码器，我们采用了与稳定扩散相同的文本 CLIP (Rombach et al, 2022)。文本 CLIP 会输出 77 个向量，每个向量都是 768 维的。而为了便于训练，我们增加了一个额外的线性层，将每个向量的维数降低到

64，得到最终的文本嵌入y_0。此外，我们基于 GPT-2 构建了文本解码器D^{text} (Radford et al, 2019)。具体来说，GPT-2 将y_0作为前缀嵌入(Mokady et al, 2021)，并自回归重建文本。此外，将参数冻结在 CLIP 中，然后训练线性层并对 GPT-2 进行微调，对输入文本进行重构，这样的重构效果良好。而我们在附录 E 中提供了更多的训练细节和重建结果。（P5）

4.2. 基于 Transformer 的联合噪声预测网络

我们根据公式(5)以及在第 4.1 节中获得的嵌入来训练联合噪声预测网络。而在 UniDiffuser 中，我们使用基于 Transformer 的骨干来处理来自不同模态的输入是很自然的。**特别地，我们采用了 U-ViT** (Bao et al, 2022a)，一种最近提出的用于条件扩散模型的 Transformer。最初的 U-ViT 的特点是将所有输入(包括数据、条件和时间步长)视为标记，并在浅层和深层之间使用长跳跃连接。

在 UniDiffuser 中，我们稍微修改了 U-ViT，将数据的两种形式及其对应的时间步视为令牌。此外，我们通过经验发现，原始 U-ViT 中的层前归一化(Xiong et al, 2020)在混合精度训练时容易产生溢出。因此，一个简单的解决方案是使用**层后归一化**(Vaswani et al, 2017)，并在连接长跳跃连接后再添加层归一化，这样就稳定了 UniDiffuser 的训练。**我们在图 4 (b)中说明了模型的主干结构图**，并在附录 D 中提供了更多细节。（P5）

5. 相关工作

多模态生成建模。许多先前关于多模态生成建模的工作可以形式化为**学习条件分布**。代表性的应用包括文本到图像的生成 (Ramesh et al., 2021; Ding et al., 2021; Ramesh et al., 2022; Nichol et al., 2022; Saharia et al., 2022; Yu et al., 2022; Gu et al., 2022; Xu et al.,2018; Rombach et al., 2022)，文本到视频的生成(Ho et al., 2022a)，文本到语音的生成(Chen et al., 2021;Popov et al., 2021)和图像的字幕(即图像到文本的生成) (Mokady et al., 2021; Chen et al., 2022)。**但这些模型都是专门为单一任务而设计的。**除了学习条件分布之外，Hu et al(2022)还旨在通过离散扩散模型学习图像和文本数据的联合分布(Gu et al, 2022)。然而，它的可伸缩性尚

未被开发。（P6）

最相关的前期工作是 Versatile Diffusion (VD) (Xu et al, 2022)，它采用的是多流架构，并在传统的多任务框架中为多生成任务进行训练，但这需要多个前馈来计算所有任务的损失，并在训练期间为不同层仔细调整梯度乘数。而相比之下，UniDiffuser 提供了一个基于训练扩散模型的、深度统一视图的优雅解决方案。**因此，UniDiffuser 更简单(只有一次训练损失)，训练更有效(每次更新只有一次向前向后)，并且可以处理更多的任务(能够执行联合采样)。**此外，在我们的实验中，UniDiffuser 在图像到文本和文本到图像生成任务中的 FID 和 CLIP 得分都优于 VD(见第 6 节)，这表明 UniDiffuser 中的时间条件策略在统计上比 VD 中的多任务策略更有效。（P6）

多模态表示学习旨在学习不同模态的特征，这些特征可以转移到下游任务中。视觉和语言预训练(VLP)是最重要的。VLP 可以采用不同的策略，如对比学习(Radford et al., 2021)，潜在数据建模(Radford et al., 2021)和多重损失的组合 (Kim et al., 2021; Li et al., 2022; Bao et al.,2022e)。此外，通常使用 Transformer 来融合两种模态。总之，这项工作意味着 Transformer 对于多模态生成建模也是有效的。

扩散模型最初是由 Sohl-Dickstein 等人(2015)提出的。但最近，Ho 等人(2020)引入了一种噪声预测公式，而 Song 等人(2021c)引入了学习扩散模型的随机微分方程公式。扩散模型能够生成高质量的图像(Dhariwal & Nichol, 2021)，音频(Chen et al., 2021; Kong et al., 2021)，视频(Ho et al., 2022b),，点云(Luo & Hu, 2021)和分子构象(Hoogeboom et al., 2022; Bao et al., 2022d)。扩散模型的其他改进包括快速采样(Song et al., 2021a; Bao et al., 2022c; Salimans & Ho, 2022;Lu et al., 2022b;c)，改进训练和采样技术等。(Nichol & Dhariwal, 2021; Song et al., 2021b;Kingma et al., 2021; Vahdat et al., 2021; Zhao et al., 2022;Bao et al., 2022b; Lu et al., 2022a; Karras et al., 2022).

6. 实验

我们将在第 6.1 节中介绍实验设置。我们在第 6.2 节中展示了 UniDiffuser 执行多个生成任务的能力，并直接将其与现有的大型模型进行了比较。我们进一步证明 UniDiffuser 能自然地支持数据变化、模态之间的阻塞吉布斯采样（blocked Gibbs sampling）(见章节 6.3)和图像之间的插值(见章节 6.4)等应用。（P7）

6.1. 设置

数据集。在稳定扩散(Rombach et al, 2022)之后，我们使用了 LAION-5B 的三个子集(Schuhmann et al, 2022)。第一个是 laion2B-en，它包含大约 2B 对图片—文本和英文字幕。第二种是 laion-high-resolution，它包含 170M 左右的图像—文本对，图像分辨率≥1024，含多语言字幕。第三个是 laion-aesthetics v2 5+，它是 laion2b-en 的一个子集，包含大约 600M 个高视觉质量的图像—文本对。在稳定扩散之后，我们将 laion-aesthetics v2 5+过滤为分辨率≥512 且估计水印概率<0.5 的图像，从而保留了大约 193M 对。此外，由于 LAION-5B 中的文本相当嘈杂，我们通过删除 url、HTML 标记、电子邮件、括号中的内容、除 's 之外的引号和除“， 。？ !”之外的符号来进一步清理 laion-aesthetics v2 5+子集中的文本。（P7）

训练和抽样。在稳定扩散之后，训练是多阶段的(Rombach et al, 2022)。在第一阶段，我们在 laion2B-en 上以 256×256 的分辨率训练 250K 步，批大小为 11264，预备步长大小为 5K。在第二阶段，我们对模型进行微调，在 laion-high-resolution 上以 512×512 分辨率训练 200K 步，批大小为 2112，预备步长为 5K。在最后一个阶段，我们从第二阶段的最后一个检查点恢复（包括模型的权重和优化器的状态），并在 laion-aesthetics v2 5+上以 512×512 分辨率训练 220K 步，批处理大小为 2112。（P7）

继 Bao 等人(2022a)之后，我们使用 AdamW 优化器(Loshchilov & Hutter, 2019)进行优化，其学习速率为 2e-4，权重衰减为 0.03，在所有阶段的运行系数为$(\beta 1, \beta 2) = (0.9, 0.9)$。然后，我们将学习率降低 10 倍，并在验证损失没有减少时继续训练。而为了提高效率，我们还进行了混合精度的训练。当 U-ViT 在 256×256

分辨率下进行训练时，我们用双线性插值的方法来插值与图像相关的位置嵌入。此外，我们在所有实验中使用 DPM-Solver (Lu et al, 2022b;c)有 50 个步骤。（P7）

基线。据我们所知，多功能扩散(VD) (Xu et al, 2022)是通用多模态生成的最直接竞争对手(详见第 5 节)。如果可能，我们在所有实验中都直接与 VD 进行比较。由于原论文中没有定量的结果，所以 VD 的结果是我们按照官方代码复制的。

评估。对于从文本到图像的生成，我们报告了 MS-COCO 验证集(Lin et al, 2014)上的 FID (Heusel et al, 2017)和 CLIP 评分(Radford et al, 2021)，它们分别测量了图像保真度和图像—文本对齐。根据文献，我们从 MS-COCO 验证集中随机抽取 10K 和 30K 提示来计算生成图像的 FID 和 CLIP 评分。而对于图像—文本生成，我们报告 CLIP 评分来衡量图像—文本对齐。具体来说，我们通过随机绘制 10K 图像来计算生成文本的分数。（P7）

6.2. 主要结果

我们首先系统地比较了最直接的基线通用扩散模型(VD)。在文本到图像和图像到文本的生成领域中，这是一个通用的生成模型。而在定量上，UniDiffuser 在所有指标和指导 CFG 尺度下的两项任务中都始终优于 VD，如图 5 和图 6 所示。（P8）

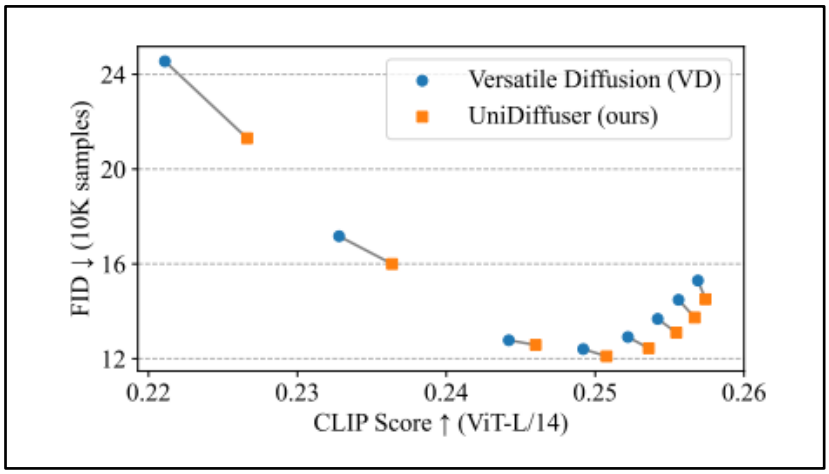

图 5：比较了 UniDiffuser 和 VD 在文本到图像生成中的作用。我们将 CFG 中相同尺度的结果连接起来。

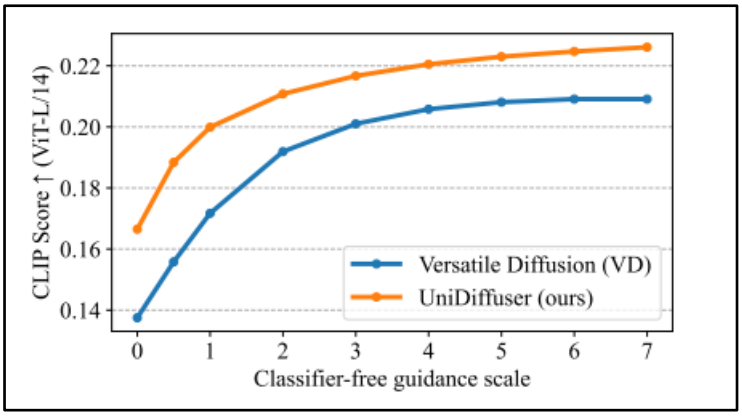

图 6：比较了 UniDiffuser 和 VD 在图像到文本生成中的作用。UniDiffuser 始终优于 VD 具有相同的 CFG 尺度(水平轴)，CLIP 评分↑(垂直轴)。

总之，经验结果证明了 UniDiffuser 与 VD 相比的有效性(除了简单性、效率和通用性之外)(参见第 5 节)。定性地说，图 7 中展示了文本到图像生成中的样本，**UniDiffuser 能比 VD 更好地对齐图像和文本**。在附录 G 中可以看到更多的结果，包括图像到文本的生成。

图 7：从文本到图像生成的 UniDiffuser 和 VD 的随机样本。UniDiffuser 生成语义正确的图像会给出代表性提示，而 VD 没有。

我们还比较了为 MS-COCO 上的零镜头 FID 而设计的文本到图像生成的定制系统在表 1 中。尽管 **UniDiffuser 设计用于处理多个生成任务**，但其在单个文本到图像生成任务上的性能与定制扩散模型(如 Stable diffusion)相当，并且优于著名的扩散模型(例如 DALL·E 2.)。（P8）

Model	FID ↓
Bespoken models	
GLIDE (Nichol et al., 2022)	12.24
Make-A-Scene (Gafni et al., 2022)	11.84
DALL·E 2 (Ramesh et al., 2022)	10.39
Stable Diffusion† (Rombach et al., 2022)	8.59
Imagen (Saharia et al., 2022)	7.27
Parti (Yu et al., 2022)	**7.23**
General-purpose models	
Versatile Diffusion† (Xu et al., 2022)	10.09
UniDiffuser (**ours**)	**9.71**

表 1：MS-COCO 验证集上的零镜头 FID。

“†”用于标记我们在正式实施时产生的结果，而其他结果均取自相应的参考文献。我们在 CFG 中报告了 UniDiffuser 和 VD 的结果，CFG 的尺度为 3，根据图 5，这是两个模型的最佳选择。

最后，我们在图 1 (a-e)中展示了联合、条件和无条件生成的示例，以展示 UniDiffuser 的一般性。更多例子见附录 A。（P8）

6.3. 数据变化和 Gibbs 抽样

UniDiffuser 自然支持图像变化和文本变化等应用程序。例如，给定一个源图像，**我们可以先进行图像到文本的生成，得到图像的描述，然后以该描述为输入进行文本到图像的生成，得到语义相似但内容不同的新图像。**在图 1 (f-g)中，我们给出了图像和文本变化的例子。此外，我们可以执行阻塞吉布斯采样（blocked Gibbs sampling），以查看图像和文本是如何通过 UniDiffuser 建模的链式条件分布来相互转换的。我们在图 1 (h)中给出了例子。而更多关于数据变化和 blocked Gibbs sampling 可以在附录 A 中找到。（P8）

6.4. 在随意的两个图像之间进行插值

UniDiffuser 还可以在两个图像之间执行插值。具体来说，我们首先进行图像—文本生成，然后通过确定性 DPM-Solver 以相同的高斯噪声作为两张图像的初始状态，获得两张图像的潜在文本嵌入。然后，通过 DPM-Solver 进行噪声注入，在给定两个潜在文本嵌入的情况下，得到一个有噪声的潜在图像嵌入。我们在潜在文本嵌入和潜在图像嵌入的噪声版本之间进行球面线性插值，以获得中间状态。最后，以文本中间状态为条件，图像中间状态为初始状态，通过 DPM-Solver 生成最终的图像。有关插值过程的形式化算法，请参阅附录 F。而我们在图 1 (i)中展示了一些例子，更多的例子可以在附录 A 中找到。（P8）

7. 结论

我们提出了 UniDiffuser，这是一个通用的多模态概率框架，基于对不同分布的扩散模型统一训练的见解。UniDiffuser 能够通过一个模型执行各种生成任务，只需对原始扩散模型进行最小的修改。而对图像—文本数据的实证结果表明，该方法的有效性与现有的大型模型相比，UniDiffuser 更为有效。UniDiffuser 还支持半监督学习和更多模式的学习，这些都是未来的工作。

但目前，我们的实现生成的文本还不是那么流畅，而这主要是因为数据有噪声。（P8）

社会影响：由于 UniDiffuser 的通用性，我们相信 UniDiffuser 可以通过生成内容来推进现实世界的应用程序。然而，值得注意的是，大规模的多模态生成模型可能会产生类似“深度造假”的后果。我们对从模型中采样的所有图像进行水印处理，并在发布代码和模型之前提供一个系统的协议来解决这个问题。（P8）