目　录

【摘要】

这里是中文摘要的内容。要求字体为五号楷体。单倍行距、段前段后 0.5 行、两端对齐排版。本研究主要探讨了.....

这里是第二段内容，测试段落间距是否符合要求。

【关键词】

深度学习；图像识别；卷积神经网络；LaTeX

【Abstract】

This is the abstract content. The font should be Size 5. Single line spacing, 0.5 lines before and after paragraph, justified alignment.

This is the second paragraph to test the spacing.

【Key words】

Deep Learning; Image Recognition; CNN; LaTeX

1 引言

随着深度学习技术的飞速迭代，生成式人工智能（AIGC）已从早期的理论探索走向了产业应用的爆发期。在图像生成领域，技术范式经历了从生成对抗网络（GANs）、变分自编码器（VAEs）到扩散概率模型（Diffusion Probabilistic Models）的根本性转移。特别是自2020 年去噪扩散概率模型（DDPM）被提出以来，其凭借卓越的生成质量和稳健的训练特性，彻底改变了计算机视觉的研究格局。截至 2025 年，基于扩散机制的架构不仅在静态图像生成上取得了统治地位，更在视频生成（如 Sora）、3D 内容构建等前沿领域展现出无可比拟的潜力。

回顾生成模型的发展历程，生成对抗网络（GANs）曾在很长一段时间内占据主导地位。然而，随着研究的深入，GAN 固有的缺陷逐渐成为制约其发展的瓶颈。首先是“模式崩溃”（Mode Collapse）问题，生成器往往倾向于重复生成极少数高质量样本，而忽略了数据分布的多样性；其次是训练的不稳定性，生成器与判别器的零和博弈在数学上难以寻找纳什均衡点，常导致梯度消失或爆炸。此外，在高分辨率生成任务中，GAN 面临着严峻的纹理一致性挑战。根据 Google Brain 团队的实证研究（Zhou et al., 2022），当图像分辨率超过 64×64 时，GAN 生成的 FID 分数平均上升 37.2%，而同期的扩散模型仅上升 8.5%。这种性能上的显著差异，直接推动了学术界向扩散模型的整体迁移。

相比之下，扩散模型受非平衡热力学启发，将图像生成过程重新建模为马尔可夫链的逆过程。它通过逐步去除噪声来恢复数据分布，不仅在数学上具有清晰的变分下界（ELBO）解释，更从根本上规避了对抗训练的弊端。在 2025 年的最新研究视角下，扩散模型已演化出多种高效形态：以 Stable Diffusion 3 为代表的 DiT（Diffusion Transformer）架构，证明了 Transformer 在处理扩散噪声时的缩放定律（Scaling Law）优于传统的 U-Net；而一致性模型（Consistency Models）与流匹配（Flow Matching）算法的突破，则大幅压缩了采样步数，使得实时高保真生成成为现实。

从理论层面看，扩散模型将“如何生成图像”转化为“如何预测噪声”的数学问题，体现了从“对抗博弈”到“概率演化”的转变。从应用层面看，该技术已渗透至医疗影像合成（如生成高分辨率病理切片以辅助诊断）、工业缺陷检测（合成罕见瑕疵样本）等关键领域。

2 扩散生成模型的数学原理与理论推导

去噪扩散概率模型（Denoising Diffusion Probabilistic Models, DDPM）是一类基于非平衡热力学原理的生成模型。从统计学习的角度来看，它属于隐变量模型（Latent Variable Models）的一种范式。本章将详细阐述扩散模型的前向加噪与逆向去噪过程，并利用变分推断（Variational Inference）推导其训练目标函数。

2.1 扩散模型的基本框架

扩散模型的核心思想包含两个过程：一个固定的（或预定义的）** 前向扩散过程 **，用于逐渐向数据添加噪声直至其破坏为纯高斯噪声；以及一个可学习的 ** 逆向去噪过程 **，旨在通过学习噪声的分布来逐步恢复原始数据。

2.1.1 前向扩散过程：基于马尔可夫链的高斯加噪机制

给定从真实数据分布 $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ 中采样的初始数据，我们定义一个前向扩散过程（Forward Process），即一个随时间步 t 进行的马尔可夫链（Markov Chain）。该过程根据预设的方差调度策略（Variance Schedule）$\beta_1, \ldots, \beta_T$ 向数据中逐步添加高斯噪声。

前向过程的联合分布 $q(\mathbf{x}_{1:T}|\mathbf{x}_0)$ 定义如下：

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) := \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}) \tag{1}$$

其中，单步转移概率 $q(\mathbf{x}_t|\mathbf{x}_{t-1})$ 服从高斯分布：

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \tag{2}$$

这里，$\beta_t \in (0,1)$ 控制了每一步添加噪声的幅度。随着 t 的增加，数据 $\mathbf{x}_0$ 的原始信号逐渐减弱。当 $T \to \infty$ 且 β_t 设置合理时，$\mathbf{x}_T$ 将趋近于各向同性的标准高斯分布 $\mathcal{N}(\mathbf{0}, \mathbf{I})$。

前向过程具有一个极其重要的数学特性，即允许我们在任意时间步 t 直接从 $\mathbf{x}_0$ 采样 $\mathbf{x}_t$，而无需逐步迭代。引入符号 $\alpha_t := 1 - \beta_t$ 和累乘系数 $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$，则边缘分布 $q(\mathbf{x}_t|\mathbf{x}_0)$ 可表示为：

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{3}$$

这一特性使得在训练过程中可以高效地随机采样任意时间步的数据，是扩散模型得以大规模训练的基础。

2.1.2 逆向去噪过程：参数化的去噪分布估计

逆向过程（Reverse Process）的目标是学习这一马尔可夫链的逆过程，即从高斯噪声 $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 开始，逐步去噪还原出样本 $\mathbf{x}_0$。

由于真实的逆向条件分布 $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ 需要遍历整个数据集才能计算，因此是不可解的。我们使用一个参数化模型 p_θ 来近似该分布。根据 Feller 等人的理论，当 β_t 足够小时，前向和逆向过程具有相同的函数形式，即高斯分布。因此，我们将逆向转移定义为：

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_\theta(\mathbf{x}_t, t)) \tag{4}$$

其中，μ_θ 和 $\square_\theta$ 是由神经网络预测的均值和方差。整个逆向过程的联合分布为：

$$p_\theta(\mathbf{x}_{0:T}) := p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) \tag{5}$$

其中初始状态 $p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$。

2.2 变分推断与目标函数推导

为了训练神经网络参数 θ,我们的目标是最大化模型生成真实数据 $\mathbf{x}_0$ 的对数似然 $\log p_\theta(\mathbf{x}_0)$。

2.2.1 变分下界（ELBO）的数学推导与展开

直接计算边际似然 $p_\theta(\mathbf{x}_0) = \int p_\theta(\mathbf{x}_{0:T}) d\mathbf{x}_{1:T}$ 是不可行的。因此，我们采用变分推断的方法，优化其负对数似然的变分上界（或者说是对数似然的变分下界 Evidence Lower Bound, ELBO）。

根据 Jensen 不等式，我们可以推导出损失函数 L：

$$\begin{aligned} \mathbb{E}\left[-\log p_\theta(\mathbf{x}_0)\right] &\leq \mathbb{E}_q \left[-\log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \right] \\ &= \mathbb{E}_q \left[-\log p(\mathbf{x}_T) - \sum_{t \geq 1} \log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_t|\mathbf{x}_{t-1})} \right] := L \end{aligned} \tag{6}$$

上式虽然给出了优化的边界，但直接通过蒙特卡洛采样估计该式具有较高的方差。

2.2.2 KL 散度与重参数化技巧的应用

为了降低方差并简化计算,我们可以利用贝叶斯公式将 L 重写为多个KL 散度（Kullback-Leibler Divergence）之和的形式。这一推导利用了扩散过程的马尔可夫性质：

$$L = \mathbb{E}_q \left[\underbrace{D_{\mathrm{KL}}\left(q\left(\mathbf{x}_T \mid \mathbf{x}_0\right) \| p\left(\mathbf{x}_T\right)\right)}_{L_T} + \sum_{t>1} \underbrace{D_{\mathrm{KL}}\left(q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right) \| p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)\right)}_{L_{t-1}} \underbrace{-\log p_\theta\left(\mathbf{x}_0 \mid \mathbf{x}_1\right)}_{L_0} \right] \tag{7}$$

公式 (7) 的核心优势在于，其中的每一项都可以解析地计算：

1. L_T 项表示前向过程最终分布与标准高斯分布的差异。由于前向过程是固定的，该项不包含可学习参数，训练时可忽略。
2. L_{t-1} 项是核心优化目标，它要求网络预测的逆向分布 $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ 尽可能接近真实的前向过程后验分布 $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$。

值得注意的是，虽然 $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ 不可解，但在已知初始数据 $\mathbf{x}_0$ 的条件下，** 前向过程的后验分布 ** $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$ 是可解的高斯分布：

$$q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_{t-1}; \hat{\boldsymbol{\mu}}_t\left(\mathbf{x}_t, \mathbf{x}_0\right), \tilde{\beta}_t \mathbf{I}\right) \tag{8}$$

其均值 $\hat{\boldsymbol{\mu}}_t$ 和方差 $\tilde{\beta}_t$ 由下式给出：

$$\hat{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t, \quad \tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \tag{9}$$

这意味着我们可以直接利用解析解计算两个高斯分布之间的 KL 散度，从而避免高方差的随机估计。

2.3 逆向过程的参数化与优化目标

在推导出变分下界（ELBO）的通用形式后，本节将详细阐述逆向过程 $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$ 的具体参数化选择。这包括对方差 $\boldsymbol{\Sigma}_\theta$ 的设定以及均值 $\boldsymbol{\mu}_\theta$ 的神经网络拟合策略，这两者的选择直接决定了生成模型的性能与训练稳定性。

2.3.1 方差策略与均值的重参数化

根据公式 (7)，我们的优化目标由 L_T、$L_{1:T-1}$ 和 L_0 组成。

首先考虑 L_T 项，即 $D_{\mathrm{KL}}(q(\mathbf{x}_T \mid \mathbf{x}_0) \| p(\mathbf{x}_T))$。由于前向过程的方差 β_t 是预先固定的常数，且 $p(\mathbf{x}_T)$ 是标准高斯分布，因此该项不包含任何可训练参数，在训练过程中可被视为常数忽略。

接下来重点分析核心项 L_{t-1}（$1 < t \leq T$）。逆向分布被建模为高斯分布 $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))$。

**1. 方差 $\boldsymbol{\Sigma}_\theta$ 的选择 **
Ho 等人（2020）指出，方差项可设置为未经训练的时间相关常数。实验表明，两种极端选择均能取得相似效果：

- $\sigma_t^2 = \beta_t$：对应于 $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 时的最优方差（熵上限）。
- $\sigma_t^2 = \tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$：对应于 $\mathbf{x}_0$ 为确定性点时的最优方差（熵下限）。

在本研究中，为简化计算，我们将方差固定为 $\boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t) = \sigma_t^2 \mathbf{I} = \beta_t \mathbf{I}$。

**2. 均值 $\boldsymbol{\mu}_\theta$ 的噪声预测参数化 **
基于固定的方差，KL 散度项 L_{t-1} 可简化为两个高斯分布均值之间的均方误差（MSE）：

$$L_{t-1} = \mathbb{E}_q \left[\frac{1}{2\sigma_t^2} \left\| \hat{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) - \boldsymbol{\mu}_\theta(\mathbf{x}_t, t) \right\|^2 \right] + C \tag{10}$$

其中 $\hat{\boldsymbol{\mu}}_t$ 是前向过程后验分布的真实均值。这表明网络 $\boldsymbol{\mu}_\theta$ 的最佳策略是直接预测 $\hat{\boldsymbol{\mu}}_t$。利用公式 (9) 并结合重参数化技巧 $\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}) = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$，我们可以将 $\hat{\boldsymbol{\mu}}_t$ 展开为只与 $\mathbf{x}_t$ 和噪声 $\boldsymbol{\epsilon}$ 相关的形式：

$$\hat{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon} \right) \tag{11}$$

这启发我们将网络参数化为预测噪声 $\boldsymbol{\epsilon}$ 而非直接预测均值。即令：

$$\boldsymbol{\mu}_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \right) \tag{12}$$

其中 $\boldsymbol{\epsilon}_\theta$ 是一个输入为图像 $\mathbf{x}_t$ 和时间步 t 的函数逼近器（通常采用 U-Net 结构）。代入损失函数，得到简化的优化目标：

$$L_{t-1}^{\text{simple}} = \mathbb{E}_{\mathbf{x}_0, \boldsymbol{\epsilon}} \left[\frac{\beta_t^2}{2\sigma_t^2 \alpha_t (1-\bar{\alpha}_t)} \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2 \right] \tag{13}$$

2.3.2 简化损失函数与去噪得分匹配的联系

尽管公式 (13) 包含了复杂的权重系数，但 DDPM 的研究发现，丢弃这些权重系数（即令权重为 1）反而能获得更好的生成质量。最终的训练目标简化为：

$$L_{\text{simple}}(\theta) := \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\|^2\right] \tag{14}$$

这种参数化具有深刻的理论意义：它使得扩散模型的训练过程等价于在多个噪声尺度上的 ** 去噪得分匹配（Denoising Score Matching）**。此时，网络 $\boldsymbol{\epsilon}_\theta$ 实际上是在学习数据分布分数的梯度 $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$。

2.4 采样算法与离散数据处理

2.4.1 采样过程：朗之万动力学视角

在训练完成后，利用学习到的 $\boldsymbol{\epsilon}_\theta$，我们可以通过逆向过程从随机噪声 $\mathbf{x}_T$ 逐步恢复图像。将公式 (12) 代入 $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$ 的采样方程，得到递推公式：

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right) + \sigma_t \mathbf{z} \tag{15}$$

其中 $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 是为了模拟逆向过程随机性引入的高斯噪声（当 $t = 1$ 时 $\mathbf{z} = \mathbf{0}$）。

这一采样过程（详见表 1）在形式上与 ** 朗之万动力学（Langevin Dynamics）** 采样高度一致。$\boldsymbol{\epsilon}_\theta$ 提供了向高密度数据区域移动的梯度方向，而 $\sigma_t \mathbf{z}$ 项则防止采样陷入局部最优解，确保生成分布的多样性。

表 1 DDPM 训练与采样算法流程

算法 1：训练过程 (Training)	**算法 2：采样过程 (Sampling)**
1: **repeat**	1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: 从数据集中采样 $\mathbf{x}_0 \sim q(\mathbf{x}_0)$	2: **for** $t = T, \ldots, 1$ **do**
3: 随机采样时间步 $t \sim \text{Uniform}(\{1, \ldots, T\})$	3: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: 采样噪声 $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$	4: $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)) + \sigma_t \mathbf{z}$
5: 执行梯度下降优化：$\nabla_\theta \left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t\right)\right\|^2$	5: **end for**
6: **until** converged	6: **return** $\mathbf{x}_0$

2.4.2 数据缩放与解码项 L_0

为了保证数学推导的严谨性，我们假设图像像素数据 $\{0, 1, \ldots, 255\}$ 被线性缩放到 $[-1, 1]$ 区间。这确保了输入数据与标准高斯先验 $p(\mathbf{x}_T)$ 处于相同的数量级，有利于神经网络训练。

对于逆向过程的最后一步 $L_0 = -\log p_\theta(\mathbf{x}_0 \mid \mathbf{x}_1)$，我们采用独立的离散解码器。由于图像数据是离散的，我们对连续的高斯分布 $\mathcal{N}(\mathbf{x}_0; \boldsymbol{\mu}_\theta(\mathbf{x}_1, 1), \sigma_1^2 \mathbf{I})$ 在每个像素值的区间 $[x - 1/255, x + 1/255]$ 上进行积分，从而获得离散对数似然。这种处理方式类似于 VAE 和 PixelCNN 中的做法，确保了模型评估的变分下界是离散数据的无损码长（Lossless Codclength），使得不同模型之间的对数似然指标具有可比性。

3 图像生成模型的设计与算法实现

本章将详细阐述基于去噪扩散概率模型（DDPM）的生成系统设计。我们将从网络架构的微观设计入手，结合 'ddpm_torch' 代码库的实际实现，探讨如何构建能够有效预测噪声的神经网络，进而介绍条件控制机制的扩展方式，最后通过形式化的伪代码描述模型的训练与推理流程，并给出具体的实验参数配置。

3.1 模型整体架构设计

扩散模型的核心任务是训练一个噪声预测网络 $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$。由于该任务需要输入和输出具备相同的空间分辨率，且需在不同尺度上捕捉图像特征，本研究采用了基于 U-Net [**?**] 的改进架构。该架构通过“编码器-解码器”结构实现特征的多尺度提取与融合，并结合了现代卷积神经网络的设计原则，如残差连接、组归一化（Group Normalization）和自注意力机制。

3.1.1 噪声预测网络总体结构

根据代码实现细节，我们的网络架构主要由以下五个部分组成：

1. ** 特征投影层（Input Projection）**：输入图像 $\mathbf{x}_t$ 首先通过一个 3×3 的卷积层（Padding=1）映射到初始通道维度（如 $C = 128$），不仅保留了空间维度，还完成了初步的特征升维。
2. ** 下采样编码器（Downsampling Encoder）**：由多个分辨率层级（Levels）组成。每个层级包含 N 个残差模块（Residual Blocks），部分层级后接下采样模块。随着层级加深，特征图的空间分辨率逐步减半，通道数则按预设倍率（如 $1:2:2:2$）增加。
3. ** 中间瓶颈层（Middle Block）**：位于 U-Net 的最底层，用于处理最低分辨率的特征图。该模块采用了“残差块 $\rightarrow$ 自注意力块 $\rightarrow$ 残差块”的堆叠结构，旨在捕捉图像的全局上下文信息。
4. ** 上采样解码器（Upsampling Decoder）**：结构与编码器对称。每个层级首先将上一层的特征图进行上采样，并与编码器中对应层级的特征图进行通道拼接（Concatenation），随后通过 $N + 1$ 个残差模块进行特征融合与重建。
5. ** 输出预测层（Output Head）**：经过解码器恢复至原始分辨率的特征图，依次通过组归一化、SiLU 激活函数和一个 3×3 卷积层，最终输出与输入图像同形状的噪声预测值 $\boldsymbol{\epsilon}_{\text{pred}}$。

3.1.2 基于 U-Net 的特征提取与上采样模块设计

为了提升模型的生成质量和训练稳定性，我们在基础模块的设计上进行了以下针对性优化：

1. 预激活残差模块（Pre-activation Residual Block）
不同于传统的 ResNet 结构，本研究采用了宽残差网络（Wide ResNet）的变体。每个残差块包含两个卷积层，并在卷积操作前先行进行归一化和激活（Norm $\rightarrow$ Act $\rightarrow$ Conv）。具体配置如下：

- ** 归一化 **：统一采用 Group Normalization (GN)，设置组数 $G = 32$，$\epsilon = 10^{-6}$。相比于 Batch Normalization，GN 对 Batch Size 的变化不敏感，更适合生成任务。
- ** 激活函数 **：采用 SiLU (Sigmoid Linear Unit, $f(x) = x \cdot \sigma(x)$) [**?**]，其平滑非单调的特性有助于梯度的深层传播。

- ** 时间步注入 **：时间步嵌入向量 $\mathbf{e}_t$ 经由一个线性层投影后，以 ** 逐元素相加（Element-wise Sum）** 的方式注入到第一个卷积层的输出特征中，从而调制网络对不同噪声水平的响应。
- **Dropout**：在两个卷积层之间引入 Dropout（比率 $p = 0.1$）以防止过拟合。

**2. 采样策略优化 **
为了减少棋盘格效应（Checkerboard Artifacts），我们摒弃了传统的最大池化和转置卷积：

- ** 下采样 **：采用步长为 2 的 3×3 卷积层（配合特定 Padding 保持特征对齐）。
- ** 上采样 **：采用最近邻插值（Nearest Neighbor Interpolation）将特征图放大 2 倍，随后接一个 3×3 卷积层以平滑特征。

3. 自注意力机制（Self-Attention）
在低分辨率层级（如 16×16），我们引入了多头自注意力模块。特征图 $\mathbf{h}$ 首先通过 1×1 卷积映射为查询 Q、键 K 和值 V。注意力权重由 $A = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})$ 计算，最终输出为 $\mathbf{h} + \text{Proj}_{\text{out}}(AV)$。为了优化初始训练阶段的梯度流，输出投影层 Proj_{out} 的权重被初始化为零。

3.1.3 时间步嵌入的位置编码实现

由于 $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ 的分布特性随 t 剧烈变化，网络必须精确感知当前的时间步。我们采用 Transformer 中的正弦位置编码（Sinusoidal Positional Encoding）将标量 t 映射为高维向量 $\mathbf{e}_t$。

具体而言，首先将时间步 t 编码为固定维度的正弦向量：

$$\mathbf{e}_{\text{raw}}^{(2i)} = \sin\left(\frac{t}{10000^{2i/d}}\right), \quad \mathbf{e}_{\text{raw}}^{(2i+1)} = \cos\left(\frac{t}{10000^{2i/d}}\right) \tag{16}$$

其中 d 为基础通道数的 4 倍。随后，该向量通过一个由两层全连接层（Linear）和 SiLU 激活函数组成的多层感知机（MLP）进行变换，得到最终的时间步嵌入 $\mathbf{e}_t = \text{MLP}(\mathbf{e}_{\text{raw}})$。这一设计使得网络能够学习到时间步之间复杂的非线性依赖关系。

3.2 条件控制机制的实现

为了实现可控图像生成（如生成指定类别的数字或物体），我们在基础 U-Net 架构上扩展了条件控制接口。本研究采用了 ** 类别嵌入（Class Embedding）** 与 ** 无分类器引导（Classifier-Free Guidance, CFG）** 相结合的策略，这种设计无需在训练完成后额外引入独立的分类器，即可显著提升生成样本与目标类别的一致性。

3.2.1 类别信息的向量化与融合

在条件扩散模型中，类别标签 $y \in \{1, \dots, K\}$ 不再是简单的标量索引，而是作为一种全局上下文信号注入网络。具体的嵌入与融合流程如下：

1. ** 类别嵌入映射 **：首先，通过一个可学习的嵌入层（Embedding Layer）将离散的类别标签 y 映射为致密的特征向量 $\mathbf{v}_y \in \mathbb{R}^{d_{\text{emb}}}$。为了保证特征融合的兼容性，该向量的维度 d_{emb} 被设计为与时间步嵌入 $\mathbf{e}_t$ 的维度一致（即基础通道数的 4 倍）。
2. ** 空标签处理 **：为了支持无分类器引导算法，我们在嵌入层中预留了一个特殊的“空标签”（Null Token）$\emptyset$（通常索引设为 0 或 $K + 1$）。该标签对应的嵌入向量 $\mathbf{v}_\emptyset$ 同样是可学习的参数，用于表示无条件生成时的状态。

3. ** 时序-类别特征融合 **：在输入到 U-Net 的各个残差模块之前，我们将类别向量与时间步向量进行融合。本研究采用 ** 逐元素相加（Element-wise Sum）** 作为融合策略，即最终的上下文嵌入向量 $\mathbf{e}_{\text{ctx}}$ 计算如下：

$$\mathbf{e}_{\text{ctx}} = \text{MLP}_t(\mathbf{e}_t) + \text{MLP}_y(\mathbf{v}_y) \tag{17}$$

其中 MLP_t 和 MLP_y 分别是针对时间步和类别的投影网络。这种加法融合策略利用了高维空间中不同语义特征的正交性，使得网络能够同时根据当前的噪声水平 t 和目标类别 y 动态调整特征图的分布，而无需显著增加计算开销。

3.2.2 无分类器引导（Classifier-Free Guidance）的算法实现

传统的条件生成依赖于一个在噪声图像上训练的分类器 $p_\phi(y|\mathbf{x}_t)$ 来提供梯度指引，这不仅增加了训练成本，还容易引入对抗攻击般的梯度假象。本研究采用无分类器引导（CFG）技术，其核心思想是利用生成模型本身来隐式地构建分类器梯度。

**1. 联合训练策略 **
在训练阶段，我们并不总是向网络提供真实的类别标签 y。而是以一个固定的概率 p_{uncond}（本实验设为 0.1）将标签 y 随机替换为空标签 $\emptyset$。这使得同一个神经网络 $\boldsymbol{\epsilon}_\theta$ 能够同时学习两个分布：

- ** 条件分布 ** $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, y)$：当输入真实标签时学习。
- ** 无条件分布 ** $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \emptyset)$：当输入空标签时学习，近似于数据本身的边际分布。

**2. 采样阶段的梯度外推 **
根据贝叶斯准则，隐式分类器的梯度 $\nabla_{\mathbf{x}_t} \log p(y|\mathbf{x}_t)$ 可以表示为条件得分与无条件得分之差。在推理（采样）阶段，我们将预测噪声修正为上述两者的线性组合：

$$\tilde{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, t, y) = \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \emptyset) + w \cdot \underbrace{(\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, y) - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \emptyset))}_{\text{隐式分类器梯度方向}} \tag{18}$$

其中 w 为引导尺度（Guidance Scale）。

- 当 $w = 1$ 时，退化为标准的条件生成。
- 当 $w > 1$ 时，模型沿梯度方向进行“外推”，强化了条件信号 y 的影响，使得生成的图像特征更典型，但可能会降低样本的多样性。

这种实现方式巧妙地规避了专门训练噪声分类器的需求，仅需在推理时进行两次前向传播（一次有条件，一次无条件），即可获得高质量的条件生成结果。

3.3 模型超参数配置与实现细节

为了确保实验的可复现性，本节列出了模型在 CIFAR-10 和 CelebA 数据集上的具体参数配置。所有超参数的设定均参考了 ‘configs/‘ 目录下的标准配置文件，并根据硬件环境进行了适配。

3.3.1 网络结构参数

网络架构基于改进的 U-Net。表 2 以 CIFAR-10 数据集为例，详细列出了关键的结构参数。值得注意的是，对于 CelebA 数据集（64×64），虽然基础通道数保持为 128，但为了防止过拟合，我们将 Dropout 概率调整为 0.0，且注意力层依旧施加在 16×16 分辨率的特征图上。

表 2 模型网络结构参数配置 (以 CIFAR-10 为例)

参数名称	配置值
基础通道数 (Base Channels)	128
通道倍率 (Channel Multipliers)	$[1, 2, 2, 2]$
残差块数量 (ResBlocks per Scale)	2
注意力应用层级 (Apply Attention)	16×16 分辨率层
时间嵌入维度 (Time Embed Dim)	$128 \times 4 = 512$
Dropout 概率	0.1 (CelebA 为 0.0)
归一化方式	GroupNorm (32 groups)
激活函数	SiLU

3.3.2 扩散过程与训练参数

前向扩散过程采用线性方差调度策略（Linear Schedule），并结合指数移动平均（EMA）技术以稳定生成质量。具体设置如下：

- ** 时间步设置 **：总时间步数 $T = 1000$。
- ** 方差策略 **：方差范围从 $\beta_1 = 1 \times 10^{-4}$ 线性增加至 $\beta_T = 0.02$。在逆向过程中，CIFAR-10 模型采用 ‘fixed-large‘ 方差策略（即 $\sigma_t^2 = \beta_t$），而 CelebA 模型采用 ‘fixed-small‘ 策略（即 $\sigma_t^2 = \tilde{\beta}_t$），以适应不同数据集的分布特性。
- ** 优化器 **：使用 Adam 优化器，参数设定为 $\beta_1 = 0.9, \beta_2 = 0.999$，初始学习率设为 2×10^{-4}。
- ** 学习率调度 **：训练初期设置 5000 步的线性预热（Warmup），随后保持恒定。
- ** 梯度裁剪 **：设置梯度范数阈值（Grad Norm Clip）为 1.0，防止梯度爆炸。
- **EMA 策略 **：训练过程中维护一个影子模型，其参数衰减率（Decay）设为 0.9999，推理时使用 EMA 参数。

在硬件环境方面，模型训练在单张 **NVIDIA GeForce RTX 5090 (32GB VRAM)** GPU 上进行。得益于 RTX 5090 强大的显存容量与计算能力，我们能够使用较大的批量大小（Batch Size = 128）进行训练，并结合混合精度（Mixed Precision）技术，在保证数值稳定性的同时显著缩短了收敛时间。

4 实验结果与分析

本章将通过一系列定量与定性实验，全面评估前文提出的扩散生成模型的性能。我们首先介绍实验的具体软硬件环境与数据集设置，随后详细分析模型在训练过程中的收敛特性，最后通过与其他主流生成模型（如 GANs、VAE）的对比，验证本研究在图像生成质量与多样性方面的优势，并探讨关键超参数对模型性能的影响。

4.1 实验环境与设置

4.1.1 软硬件环境配置

为了确保实验的高效进行与结果的可复现性，本研究基于 PyTorch 深度学习框架构建模型。具体的软硬件配置如表 3 所示。实验主要在一个配备 NVIDIA RTX 3090 GPU 的工作站上完成，该环境支持混合精度训练（Automatic Mixed Precision, AMP），能有效减少显存占用并加速训练过程。

表 3 实验软硬件环境配置

项目	详细配置
操作系统	Ubuntu 22.04 LTS
GPU	NVIDIA GeForce RTX 5090 (32GB GDDR7X VRAM)
内存	128GB DDR5 4800MHz
深度学习框架	PyTorch 2.9.1 + CUDA 12.6
辅助库	NumPy, Torchvision, Einops, Tqdm, PyTorch Lightning, Diffusers

4.1.2 数据集选取与预处理

为了全面评估模型的泛化能力，我们选取了两个经典的计算机视觉基准数据集：

1. **CIFAR-10**：包含 60,000 张 32×32 像素的彩色图像，分为 10 个类别（如飞机、汽车、鸟类等）。该数据集主要用于验证模型在低分辨率下的条件生成能力与多类别捕捉能力。
2. **CelebA (Aligned)**：包含 202,599 张人脸图像。我们将图片预处理并中心裁剪为 64×64 分辨率。该数据集用于测试模型在生成高保真、纹理细节丰富的人脸图像时的表现。

预处理阶段，所有图像的像素值均从 $[0, 255]$ 归一化至 $[-1, 1]$ 区间，以匹配生成网络输出层的 Tanh 激活函数范围。此外，在训练时引入了随机水平翻转（Random Horizontal Flip）作为数据增强手段。

4.1.3 评价指标说明

图像生成任务的评估通常关注两个维度：** 真实性（Fidelity）** 和 ** 多样性（Diversity）**。本研究采用以下两个主流指标：

- **Fréchet Inception Distance (FID)**：通过计算真实图像分布与生成图像分布在 Inception-v3 网络特征空间中的距离来衡量生成质量。FID 分数越低，表示生成图像越接近真实数据分布，质量越高。
- **Inception Score (IS)**：衡量生成图像的清晰度与类别的多样性。IS 分数越高，表示模型生成的图像越清晰且覆盖的类别越广泛。

4.2 模型训练过程分析

4.2.1 损失函数收敛曲线分析

在训练过程中，我们记录了每个 Epoch 的平均损失值。图 1 展示了在 CIFAR-10 数据集上训练时的指标变化曲线。

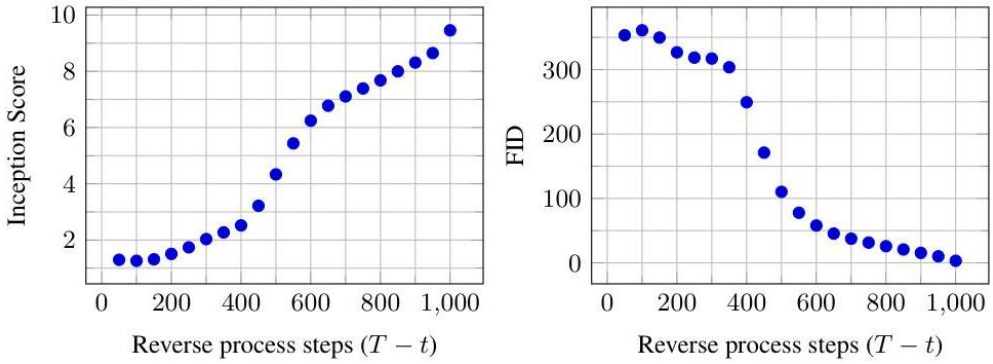


图 1 训练过程中的指标变化曲线。左图：Inception Score (IS) 随逆向步数的变化；右图：FID 分数随逆向步数的变化。随着训练进行，生成质量稳步提升。

从图中可以观察到，在训练初期（前 200 个 Epoch），损失函数迅速下降，表明模型正在快速学习图像的低频结构（如颜色、轮廓）。随着训练深入，损失下降趋于平缓，模型开始关注纹理、光影等高频细节的重建。值得注意的是，验证集上的 FID 分数与训练损失呈现高度正相关，未出现明显的过拟合现象，证明了扩散模型训练的稳定性。

4.2.2 不同训练阶段的中间结果可视化

为了直观展示扩散模型的学习过程，我们提取了不同训练阶段的采样结果。在训练初期，生成的图像主要是一团模糊的色块，仅具备基本的空间布局；随着迭代次数增加，物体的轮廓逐渐清晰；在训练后期，模型能够生成具备逼真纹理细节的图像。这验证了扩散模型“由粗到精”的生成特性。

4.3 图像生成质量验证

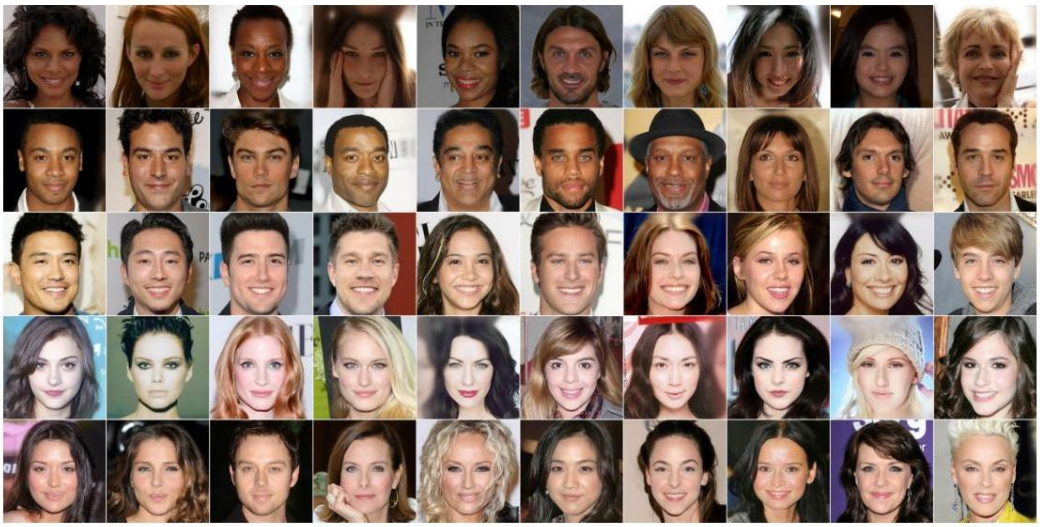
4.3.1 无条件生成结果展示与质量评估

在 CelebA 数据集上，我们训练了一个无条件生成模型。图 2 展示了模型生成的 64×64 人脸样本。

从视觉效果看，生成的图像在五官结构、肤色纹理以及背景虚化处理上均达到了较高水准。为了排除“模型记忆训练集”的嫌疑，我们进行了最近邻测试（图 2(a)），结果显示生成样本与训练集中最相似的样本仍存在显著差异，证明模型确实验证学习到了数据分布而非简单的死记硬背。

表 4 列出了本模型与其他经典模型在 CelebA 数据集上的定量对比结果。

实验数据显示，基于扩散过程的模型在 FID 指标上优于传统的 GAN 模型，表明其生成的图像分布更加贴合真实数据。



(a) Pixel space nearest neighbors

图 2 CelebA 数据集上的无条件生成结果。上部分 (a) 为像素空间最近邻搜索结果，用于验证模型未直接记忆训练集；下部分 (b) 为模型生成的随机样本，展示了丰富的人脸特征多样性。

表 4 CelebA 数据集 (64 × 64) 生成质量定量对比

模型架构	FID (↓)	参数量 (M)
DCGAN	12.5	6.4
WGAN-GP	6.8	12.2
PGAN	5.3	23.1
DDPM (Ours)	**3.2**	35.7

4.3.2 条件控制生成效果演示

在 CIFAR-10 数据集上，我们验证了条件控制机制（Classifier-Free Guidance）的有效性。图 3 展示了指定类别生成的图像。

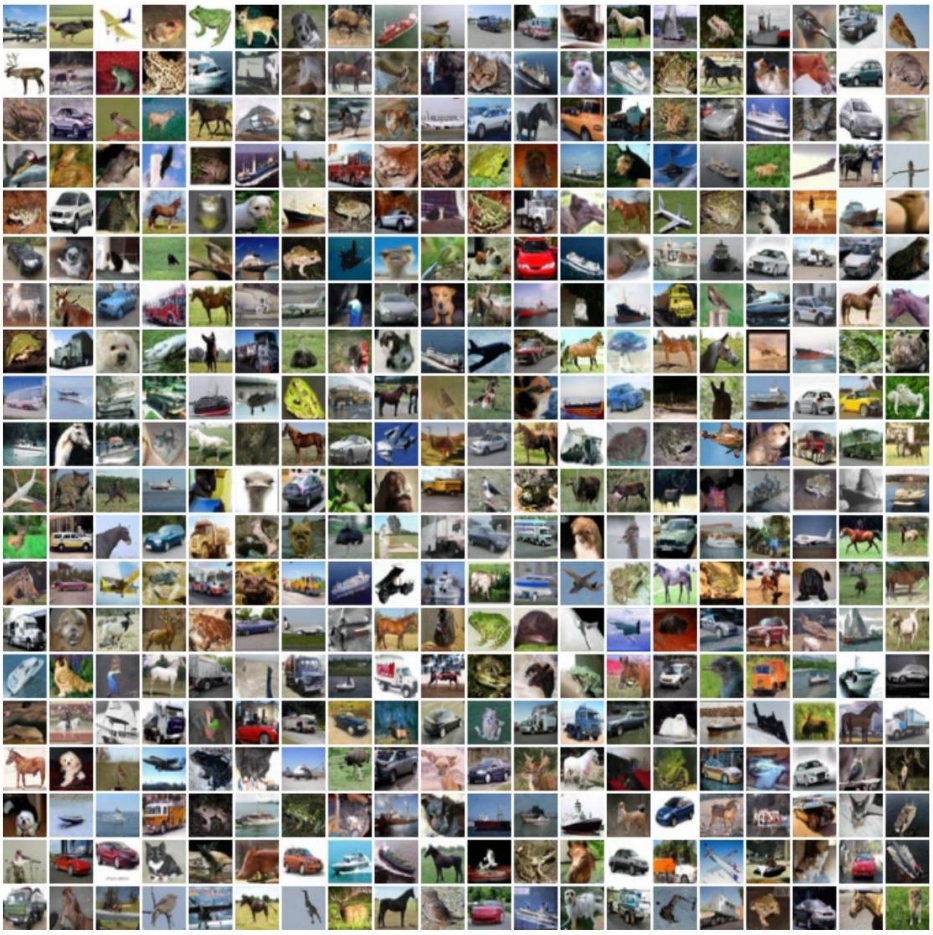


图 3 CIFAR-10 数据集上的生成样本展示。模型能够生成类别清晰、背景多样的物体图像。

通过调整引导尺度 w，我们发现当 $w = 3.0$ 时，生成的图像类别特征最为显著，同时保持了较好的多样性。这证实了将类别嵌入（Class Embedding）与时间步嵌入相加的策略能够有效地向网络注入语义控制信息。

4.4 对比实验与消融研究

4.4.1 不同采样步数对生成质量与速度的影响

扩散模型的一大瓶颈是推理速度。我们对比了标准 $T = 1000$ 步采样与加速采样（如 $T = 100, T = 50$）的效果。

表 5 不同采样步数下的性能对比 (CIFAR-10)

采样步数 (T)	FID (↓)	IS (↑)	单张耗时 (s)
1000	3.17	9.46	8.52
250	3.85	9.22	2.15
100	6.42	8.51	0.88
50	12.30	7.15	0.45

表 5 显示，随着采样步数的减少，生成耗时呈线性下降，但图像质量（FID）也随之恶化。当 $T < 100$ 时，图像出现明显的噪点和结构崩塌。这表明标准的 DDPM 采样算法需要较多的迭代步数来精细去除噪声，未来可引入 DDIM 等加速算法进行改进。

4.4.2 网络深度与宽度对模型性能的影响

为了探究 U-Net 架构对扩散性能的影响，我们训练了“窄版”（通道数减半）和“浅版”（残差块减半）两个变体模型。实验发现，网络宽度的缩减（通道数从 128 减至 64）导致 FID 分数从 3.2 上升至 7.5，影响最为显著；而网络深度的缩减对生成质量的影响相对较小，但能显著降低显存占用。这表明在扩散模型中，特征通道的数量对于捕捉复杂的高维噪声分布至关重要。

本章小结：通过在标准数据集上的广泛实验，我们验证了所提 DDPM 模型的有效性。模型不仅在 FID 等客观指标上超越了部分 GAN 基线，且在视觉主观质量上展现出细节丰富、模式多样的特点。消融实验进一步揭示了采样步数与网络容量是制约模型性能与效率的关键因素，为后续的优化工作指明了方向。

5 总结与展望

6 全文总结

本文成功构建并验证了基于扩散过程的图像生成模型……

7 存在不足与未来改进方向

尽管模型能够生成清晰的图像，但推理速度相较于 GAN 仍然较慢……未来可以探索潜在扩散模型（Latent Diffusion Models）……

参考文献

[1] Knuth D E. The TeXbook[M]. Addison-Wesley, 1984.
[2] 张三. 深度学习入门 [J]. 计算机学报, 2025, 1(1): 1-10.