Computer Science and Application 计算机科学与应用, 2025, 15(10), 126-136
Published Online October 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.1510255

Hans汉斯

基于扩散Transformer的复杂图像加速推理改进研究

兰慕辰，孙杳如

同济大学计算机科学与技术学院，上海

收稿日期：2025年9月11日；录用日期：2025年10月13日；发布日期：2025年10月22日

摘　要

文生图模型的相关研究发展迅速，使用扩散模型的Transformer架构更是其中的主流。面对当今文生图模型在复杂图像处理和提高推理速度的瓶颈时，本文参考扩散Transformer提出了一种基于图像空间复杂度的自适应生成策略。本文设计了一种结合纹理和结构信息的图像空间复杂度计算方法，并将其作为生成模型的输入，动态分析生成图像的复杂结构。基于该复杂度指标，进一步提出了分块自适应生成策略，依据子图像空间复杂度的高低来调节推理模型深度的深浅，从而在保证图像质量的同时，显著提升生成效率。我们的模型在ImageNet 256 × 256和MSCOCO 2017数据集上进行了实验验证，其在FID、sFID和IS指标上均优于主流图像生成模型，同时推理速度明显加快。生成的结果显示，其不仅能复原各种潜在的细节信息，也能有效提高模型迭代速度，证明了该方法在复杂图像生成与加速推理上的有效性和可行性。

关键词

扩散模型，空间复杂度，生成式模型

Research on Improved Complex Image Acceleration Inference Based on Diffusion Transformer

Muchen Lan, Yaoru Sun

School of Computer Science and Technology, Tongji University, Shanghai

Received: September 11, 2025; accepted: October 13, 2025; published: October 22, 2025

Abstract

Research on image-generating models has advanced rapidly, with the Transformer architecture using

文章引用: 兰慕辰, 孙杳如. 基于扩散 Transformer 的复杂图像加速推理改进研究[J]. 计算机科学与应用, 2025, 15(10): 126-136. DOI: 10.12677/csa.2025.1510255

diffusion models becoming the mainstream. Addressing the bottlenecks faced by current image-generating models in processing complex images and improving inference speed, this paper proposes an adaptive generation strategy based on image spatial complexity, drawing inspiration from the diffusion Transformer. We devise a method to calculate image spatial complexity by incorporating texture and structural information, using this as input to a generative model to dynamically analyze the complex structure of the generated image. Based on this complexity metric, we further propose a block-wise adaptive generation strategy that adjusts the depth of the inference model based on the spatial complexity of the sub-images, significantly improving generation efficiency while maintaining image quality. Our model is experimentally validated on the ImageNet 256 × 256 and MSCOCO 2017 datasets, outperforming mainstream image generation models in terms of FID, sFID, and IS metrics, while significantly accelerating inference speed. The generated results demonstrate that it not only recovers a wide range of underlying details, but also effectively improves model iteration speed, demonstrating the effectiveness and feasibility of this approach for complex image generation and accelerated inference.

Keywords

Diffusion Model, Spatial Complexity, Generative Model

1. 引言

文生图是当今生成式人工智能的重要研究和应用方向，其从文本到图像的过程，作为人类语言与生成模型之间的桥梁，该技术可以运用到包括数字人[1]、图像编辑[2]和计算机辅助设计[3]-[8]等的诸多方面。但是现有的模型训练存在进退维谷的困难，一方面多样化的场景要求更多小样本的数据集，另一方面小样本数据集难以支撑模型训练出好的结果。面对这样的难题，潜在扩散模型(LDM)，如稳定扩散[7]，在高分辨率文本生成图像任务中表现突出。究其原理，核心在于通过迭代的反向采样过程逐步去噪，逐步生成符合文本条件的高质量图像。但是这种迭代过程带来了明显的缺点：生成速度较慢，这限制了模型实时应用的可能性。针对这种问题，研究者提出了多种加速方法。例如改进 ODE 求解器，缩减采样步骤至 10~20 步[9]-[11]；另一种方法从模型本身下手，通过提炼将预训练的扩散模型压缩为仅需少量推理步骤的形式[12] [13]，Meng 等人[13]提出的两阶段提炼方法在提升无分类器引导模型采样效率方面效果显著。更进一步地，Song 等人[14]提出一致性模型作为替代方案，通过学习 ODE 轨迹上的一致性映射实现单步生成来避免计算密集的迭代过程。但这种方法仍局限于固定的像素空间图像生成，不适用于高分辨率任务；同时，条件扩散模型与无分类器指导的方法并没有完全被开发，相关的方法很难直接应用于文生图过程。LaVin-DiT 模型[15]通过引入一个时空变分自编码器和联合扩散变换器有效提升了模型的性能，但是大量增加了模型的计算量；Jia 等人[16]的研究使用创新的两阶段框架来提升生成质量，Fang 等人[17]的工作使用剪枝来优化参数量。

另一方面，文生图最重要的一步就是图像生成，为此评估图像的质量尤为重要。图像质量评估(IQA)旨在量化人类对图像质量的感知，是视觉研究中的一项基本课题，该任务在图像的恢复、压缩和渲染等应用中有很强的实用性。均方误差(MSE)一直是信号保真度和质量的标准参考指标，在算法开发中发挥了基础性作用[18] [19]。在此基础上，IQA 方法能够更准确地反映人类感知[20]-[23]，其核心的结构相似性(SSIM)指数[18]成为事实上的标准。但是这类方法依赖于图像的精确配准，对相同纹理图像的局部差异高

度敏感，具体体现在同一种纹理的不同切片对于人和机器的感受是不相同的。由于纹理表面在摄影图像中十分常见，因此开发能够与人类感知保持一致的 IQA 指标尤为重要。比如说设计一种能够统计合成纹理区域、而非逐像素重建的压缩引擎[24] [25]。这类度量方法不仅能更好地评估图像质量，还能反过来推进图像处理技术的进一步发展。

针对于当前的文字转图像生成模型，模型推理速度慢和复杂图像生成质量不高的问题，我们设计将图像质量评估以计算图像空间复杂度的方式作为图像生成模型的一个输入，自适应调整图像生成模型的深度，从而加快模型的推理，提高模型生成图像的质量。本文的主要贡献如下：

1) 本文设计了一种用于评估图像空间复杂度的方法，该方法的计算过程综合考虑了一张图像本身的纹理和结构关系，从而定量表示一张图像具体的复杂度，耦合进入图像生成模型当中，动态调控模型的生成过程；

2) 本文设计了一种根据图像复杂度从而自适应调整模型推理过程深度的生成策略，该策略在不过多更改原始模型结构的情况下，通过减少部分模型推理过程的层数，进而达到加快模型推理的效果；

3) 本文将多个主流的图像生成模型在 ImageNet 256 × 256 和 MSCOCO 2017 数据集上的 FID 成绩进行测定与横向比较，客观反映了模型的真实表现。

2. 改进复杂图像生成扩散 Transformer 模型

本文在原本 DiT 模型[26]的基础上修改了原有的注意力结构，设计了一种根据图像空间复杂度自适应更改模型深度的图像生成方法，引入针对于图像复杂部分判断标准，指导模型的图像生成过程，在 DiT 模型复杂图像生成过程中取得更好的结果。

2.1. 图像生成模型内部结构

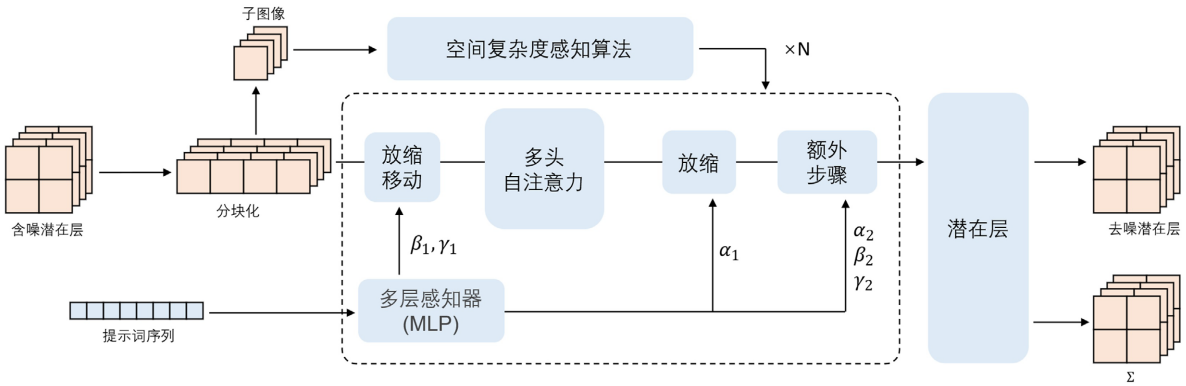

Figure 1. Schematic diagram of the denoising process of the image generation model
图 1. 图像生成模型去噪过程的结构示意图

图 1 中所展示的是大量扩散 Transformer 降噪步骤中一步的全部过程，通常降噪步长(steps)为 1000。对于上一层的输入，也就是含噪声的潜在层(noised latent)，首先将图像统一不同的通道进行分块(patchify)，分块之后的图像序列(sequence)作为变体 Transformer 模型的图像输入，提示词(caption)序列作为条件输入，分块化之后的图像序列进行放缩等一系列图像操作，期间通过计算子图像的空间复杂度而自适应调整模型的深度。最后输出的去噪潜在层(denoised latent)作为下一步的输入，另外输出对角协方差预测来计算损失。

需要指出的是，模型在一些结构上依然参考 DiT 模型的设计，比如说分块尺寸 p，在本次实验中依

然选择 $p=2,4,8$ 的情况，图像序列的长度为 p^2。模型的尺寸和深度分为 S、B、L 和 XL，模型的深度依次为 12、12、24、28，我们之后为模型处理不同的子图也采用这样的深度关系。

在模型的一些其他细节如上下文条件，我们只需将条件输入向量 t 和 c 嵌入，作为两个额外的令牌(tokens)附加到输入序列中，用图像令牌相同的处理方式进行处理，就可以取得不错的效果。类似于 ViT 中的 CLS 令牌，用于模型的分类过程，它允许我们使用标准 ViT 模块而无需修改。在模块的设计中可以在最后一个模块之后删除条件令牌，然而这种方法为模型隐性的性能开销几乎可以忽略不计。

自适应层范数(adaLN)模块沿用 DiT 的选择方案，鉴于自适应归一化层[27]在 GAN[28] [29]和以 U 型网络为骨干的扩散模型[30]中的广泛应用，我们探索用自适应层范数(adaLN)替换 Transformer 模块中的标准层范数层。我们不是直接学习维度上的缩放和平移参数和，而是根据 t 和 c 的嵌入向量之和对它们进行回归。adaLN 对于模型性能的影响是其他多种方案中最小的，因此计算效率最高。它也是唯一一种被限制为对所有令牌都应用相同函数的条件机制。

2.2. 图像纹理复杂度

Zhang 等人[31]的研究提出一种计算图像纹理复杂度的方法，该方法是一个基于深度神经网络的图像相似度度量方法，将图像输入到一个预训练神经网络，提取多层的中间特征，然后计算这些特征图之间的“距离”来衡量图像感知差异。

设两幅图像为 x 和 x_0，它们通过某个预训练神经网络 ϕ 的第 l 层提取出的特征图为 $\phi^l(x)\in R^{H_l\times W_l\times C_l}$，其中 $H_l\times W_l\times C_l$ 分别是特征图的高、宽和通道数。

对于每一层输入的图像作通道归一化处理：

$$\widehat{y^l}=\frac{\phi^l(x)}{\left\|\phi^l(x)\right\|_2+\epsilon},\ \phi^l(x)\in R^{H_l\times W_l} \tag{1}$$

其中，$\left|\phi^l(x)\right|_2$ 表示 $\phi^l(x)$ 的 L2 范数，对于 C_l 个不同的通道分别计算。ϵ 为一微量正值，保证计算的稳定并且避免出现除 0 的情况。归一化后，LPIPS 计算的不是原始特征的欧几里得距离，而是单位向量之间的加权 L2 差异。

对于原始图 x 和比对图 x_0 分别计算出的归一化特征图 $\widehat{y^l}$ 和 $\widehat{y_0^l}$，计算逐点的 L2 差异，并用通道权重 $w^l\in R^{C_l}$ 进行加权求和：

$$d^l(x,x_0)=\frac{1}{H_lW_l}\sum_{h=1}^{H_l}\sum_{w=1}^{W_l}\sum_{c=1}^{C_l}w_c^l\left(\widehat{y_{hwc}^l}-\widehat{y_{0,hwc}^l}\right)^2 \tag{2}$$

其中，$\widehat{y_{hwc}^l}$ 和 $\widehat{y_{0,hwc}^l}$ 分别表示原始图和对比图在某一个具体像素点的特征值，加权求和后的 $d^l(x,x_0)$ 表示第 l 个特征层上原始图和对比图的距离，将所有特征层叠加得到 LPIPS 距离：

$$\mathrm{LPIPS}(x,x_0)=\sum_l d^l(x,x_0) \tag{3}$$

其中，层 l 通常取神经网络中多层模型，在这里我们选用 VGG 和 AlexNet 模型。

2.3. 图像空间复杂度

Ding 等人[32]在研究过程中注意到了计算复杂度需要同时关注结构相似度和纹理相似度的感知图像质量评估指标，使得计算结果不仅可以反映全局结构的一致性，还可以反映局部细节和纹理模式的差异。我们在此基础上设计图像空间复杂度的计算方案。

同样是原始图 x 和比对图 x_0，有 $\phi^l(x),\phi^l(x_0)\in R^{H_l\times W_l\times C_l}$，对于每个位置 (h,w) 的特征向量上，先做 L2 归一化：

$$\widehat{y_{h,w}}=\frac{y_{h.w}}{\left\|y_{h.w}\right\|_2+\epsilon} \tag{4}$$

使用余弦相似度衡量两个向量之间的结构相似度：

$$struct\_sim^l=\frac{1}{H_lW_l}\sum_{h,w}\left(\left(\widehat{y_{hw}^l}\times\widehat{y_{0,hw}^{\ l}}\right)\right) \tag{5}$$

和先前过程中的推导类似，$\widehat{y_{hw}^l}$ 和 $\widehat{y_{0,hw}^{\ l}}$ 分别表示原始图 x 和比对图 x_0 在具体位置 (h,w) 的通道向量，$struct\_sim^l$ 为第 l 层的结构相似度矩阵。

使用第 l 层特征图在空间上的均值和标准差来计算图像的纹理相似度，具体定义为：

$$\mu^l(x)=\frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}\phi_{h,w}^l(x),\sigma^l(x)=\sqrt{\frac{1}{HW}\sum_{(h,w)\in R^{H_l\times W_l}}\left(\phi_{h,w}^l(x)-\mu^l(x)\right)^2} \tag{6}$$

给定具体的位置 (h,w) 分别计算该位置特征图的均值和标准差得到 $\mu^l(x)$ 和 $\sigma^l(x)$，原始图和对比图各自算一次，就得到 $\mu^l(x),\sigma^l(x),\mu^l(x_0),\sigma^l(x_0)$ 共 4 个矩阵。接下来，根据计算的均值和标准差计算通道级的亮度/对比度相似度项，具体定义为：

$$L^l=\frac{2\mu^l(x)\cdot\mu^l(x_0)+C_1}{\mu^l(x)+\mu^l(x_0)+C_1},C^l=\frac{2\sigma^l(x)\cdot\sigma^l(x_0)+C_2}{\sigma^l(x)+\sigma^l(x_0)+C_2} \tag{7}$$

其中，C_1,C_2 是两个大于 0 的小常数，其作用在于避免分母为零的情况并控制灵敏度，与特征的激活尺度匹配，在实验中值为 1e−6，L^l 和 C^l 的计算类似于调和平均数。纹理相似度的定义为亮度和对比度相似度的乘积：

$$text\_sim^l=L^l\cdot C^l \tag{8}$$

这一步计算的作用在于融合结构相似度和纹理相似度，计算得出的对比图像距离为：

$$DISTS(x,x_0)=\sum_l\alpha_l\cdot\left(1-struct\_sim^l\right)+\beta_l\cdot\left(1-text\_sim^l\right) \tag{9}$$

该距离为所有模型所有层的加权和，其中 α_l,β_l 是可学习参数。

2.4. 描述图像的复杂度

值得注意的是，上述的图像算法本质上是计算两张图像 x 和 x_0 的相似性，为了描述一张图像本身的复杂度，我们需要对对比图 x_0 做一些操作。最直接的方法是我们使用一张特定的图像作为基准，比如说令 $x_0=0_{H\times W\times C}$，即一张 $H\times W\times C$ 尺寸的纯黑图像，对于所有要参与复杂度计算的输入 x 都能得到一个稳定唯一的复杂度计算结果。但是这样会导致一些纹理复杂度不高的图像，比如说 $x=1_{H\times W\times C}$ 的纯白图像的计算结果失真。针对这种情况，我们选用高斯模糊的方案，对于图像的每个通道，有如下操作

$$x_c=x_{0,c}*G=\sum_{u=-k}^{k}\sum_{v=-k}^{k}x(h-u,w-v)G(u,v) \tag{10}$$

$$x=\sum_C x_c,x_0=\sum_C x_{0,c} \tag{11}$$

其中，G 为卷积核，k 是卷积核半径，在本次实验中我们取 $k=15$。经过高斯核模糊后的图像与原始图像进行对比，计算得出高斯模糊下的图像复杂度：

$$Gau\_Complexity(x,x_0)=\sum_l\alpha_l\cdot\left(1-struct\_sim^l(x,x_0)\right)+\beta_l\cdot\left(1-text\_sim^l(x,x_0)\right) \tag{12}$$

该公式表明所计算的图像复杂度是由原始图像与高斯核进行处理的图像，对比图像距离计算而得出的唯一指标。高斯核处理过的图像失去了一些细节和边缘信息，图像信息量减少。处理前后的图像对比，可以量化失去的信息量，从而衡量原本图像的空间复杂度。

2.5. 自适应分块生成

图 1 所示的架构中，一张噪声图需要经过分块转化成 p^2 个子图像，子图像整体作为图像序列送入 Transformer 模型之中。为了减少模型推理的计算过程、加快模型推理速度，我们设计通过计算子图像的空间复杂度而调整图像生成模型的推理深度的自适应策略。首先根据数据集的训练图像，将目标生成分为低、中、高三个档次的图像复杂度，对于低复杂度的子图像，使用最少的推理模型深度，即 $depth=12$，反之亦然，*depth* 分别设置为 24 和 28，延续 DiT 模型的设计方案。

值得一提的是，针对于目标图像的划分标准与图像数据集本身相关，在计算的过程中需要对训练数据的空间复杂度计算结果进行统计，进而形成划分标准。其次，分块的结果与 *depth* 的设定无关，每个分块所完成的操作相同，p 的不同设置可能会导致模型的不同深度，同一张图像的相同 p 设定则会出现相同的模型深度。

3. 模型训练

3.1. 数据集

我们在实验中选取了 ImageNet 256 × 256 [33]作为训练数据集，并使用 MSCOCO 2017 [34]进行模型验证。ImageNet 包含约 128 万张图像，覆盖 1000 个类别。由于该数据集内容简单、结构精巧并且有众多类别，最常用来进行图像生成模型的训练。MSCOCO 数据集涵盖 80 类常见物体，并提供了物体检测框、分割掩码及文本描述信息。该数据集一张图像的目标物体较多、内容复杂，通常不直接用于图像生成模型的研究工作，但其包含丰富的提示词，在这里我们将这个数据集作为模型的技术验证。

为了验证本研究对于复杂图像的模型训练与推理工作，我们人为地将数据集分割成三个部分：高复杂度、中复杂度、低复杂度。三种图像的比例设定为 0.25、0.5、0.25，因此每个数据集将空间复杂度排名 0.25、0.75 的计算数值作为划分点。

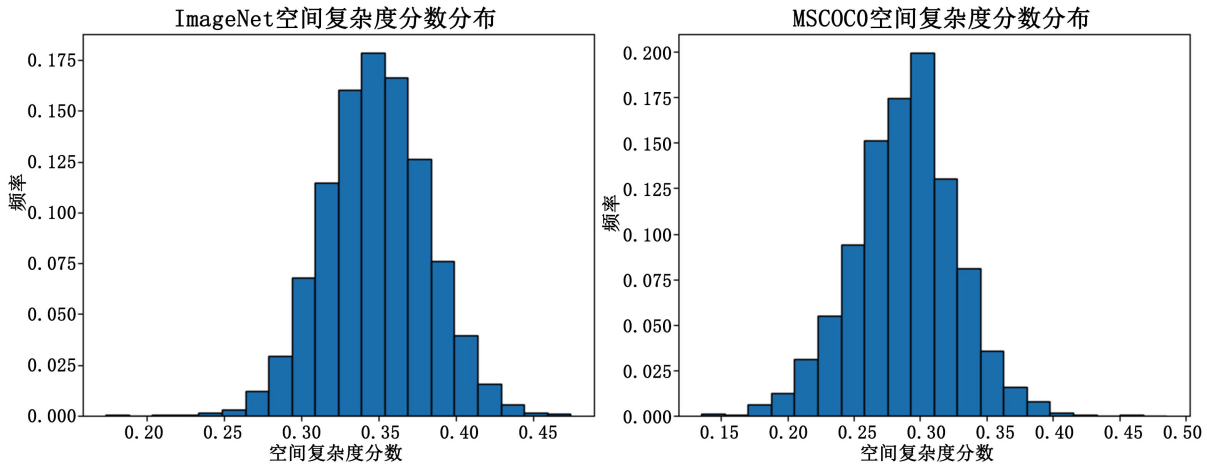

Figure 2. Frequency histograms of spatial complexity scores for ImageNet 256 × 256 and MSCOCO 2017. Each block represents the frequency of images in that score range

图 2. ImageNet 256 × 256 和 MSCOCO 2017 两个数据集的空间复杂度分数频率直方图。每一个分块代表在该分数区间下的图像频次

从图 2 可以看出，两个数据集的训练集经过上述空间复杂度算法的计算，结果分布都呈现正态分布：其分布平均值分别为 0.3481 和 0.2885，标准差分别为 0.0332 和 0.0387。根据 0.25、0.75 的分布线进行划分，ImageNet 数据集下的分数线为 0.3260 和 0.3704，MSCOCO 数据集下的分数线为 0.2642 和 0.3125。

从两个数据集的分布来看，分布的标准差基本相同，ImageNet 的分布比 MSCOCO 的分布均值更大，原因在于 ImageNet 的图像针对于目标物特意进行裁剪和缩放，图像内容单一，而 MSCOCO 的每张图像有更多的目标，图像内容更为多样。在本次实验中，MSCOCO 数据集仅作为模型的验证及性能测试。

3.2. 训练设定

在主模型框架上，我们沿用 DiT 的设计，包括模块数 $N1+N2$ 、令牌数(token)和通道数(channel)。由于 DiT 在使用较小块尺寸时展现出更强的合成性能，我们也默认使用块尺寸 $p=2$，与 LatentDiffusion 和 DiT 相同，默认采用稳定扩散提供的固定 VAE1 对图像/潜在令牌进行编码/解码。VAE 编码器的下采样率为 1/8，特征通道维度为 4，例如，大小为 $256\times256\times3$ 的图像被编码为大小为 $32\times32\times4$ 的潜在嵌入。

依照前人的工作，所有模型均使用学习率为 3e−4 的 AdamW [35]优化器进行训练，批量大小为 256，且在 ImageNet 上训练时不进行权重衰减，图像分辨率为 256×256。我们将掩码比率设置为 0.3，$N2=2$。遵循 DiT 中的训练设置，我们将训练中的最大步数设置为 1000，并使用范围在 10^{-4} 到 2×10^{-2} 之间的线性方差调度。其他设置区域也与 DiT 对齐。

3.3. 测试基线

本次实验使用到基线标准为 ImageNet 256 × 256 的 FID、sFID 和 IS 指标，以及 MSCOCO 2017 的 FID 指标。其中，FID 是利用 Inception v3 网络提取生成图像和真实图像在某一层的特征，把它们分别视作高斯分布，然后计算两个高斯分布的 Fréchet 距离，其结果越低，说明模型生成图像的性能越好，该指标兼顾了图像质量和多样性；sFID 是为了解决 FID 算法复杂度较高、样本数不足时不稳定的问题，把高维特征投影到多个一维方向，再计算一维分布之间的 Wasserstein 距离，数值越小，说明模型的性能越好；IS 同样基于 Inception v3，直接评估生成图像的类别分布而不是比较真实和生成分布，数值越大，说明模型生成的丰富度越大。

4. 实验结果

4.1. 模型对比实验

Table 1. Experimental results of mainstream image generation models on the ImageNet 256 × 256 and MSCOCO 2017 datasets, including FID, sFID, and IS metrics. Some FID results are calculated during this experiment
表 1. 主流图像生成模型在 ImageNet 256 × 256 和 MSCOCO 2017 两个数据集上的实验结果，包括 FID、sFID 和 IS 指标。部分 FID 结果为本次实验过程中计算的结果

方法	ImageNet 256 × 256			MSCOCO 2017
	FID ↓	sFID ↓	IS ↑	FID ↓
VQVAE	31.11	-	-	39.17
VQGAN	15.78	78.3	-	30.01
BigGAN-deep	6.95	7.36	**171.4**	9.94
DAE-GAN	21.32	-	-	28.12
DM-GAN	19.40	-	-	32.64

续表

DiT-XL/2	9.62	6.85	121.50	22.08
U-ViT	3.40	-	-	5.48
D^2iT-L	5.74	-	156.29	9.54
TinyDiT-D7	5.87	5.43	166.91	6.77
Ours	**2.48**	**3.93**	143.02	**5.05**

表 1 所展示的实验结果表明，我们所提出的复杂图像生成策略在 ImageNet 256 × 256 和 MSCOCO 2017 数据集上都有着优异的表现，包括在 ImageNet 256 × 256 常用的 FID、sFID 和 IS 指标，以及 MS-COCO 2017 常用的 FID 指标。我们的生成模型方案在两个数据集上的 FID 结果分别是 2.48 和 5.05，均优于现在的主流生成模型，在 ImageNet 256 × 256 数据集上的 sFID 成绩是 3.93，同样是优于主流的图像生成模型，该数据集上的 IS 得分为 143.02，大致相当于主流的中等规模参数量的模型，与各个模型最大参数量的最好结果仍有差距，其原因在于我们所设计的方案参考了 DiT 模型的中等规模参数量的设计方案，并没有完全发挥模型的全部生成性能，另一方面，我们的方案以加速推理为主，更大的模型深度和更少的分块策略会阻碍加速推理的效果实现。尽管如此，我们的模型还是在大量的实验中证明了其有效性和可靠性，在不同的数据集上的结果证明了该模型的泛化能力。

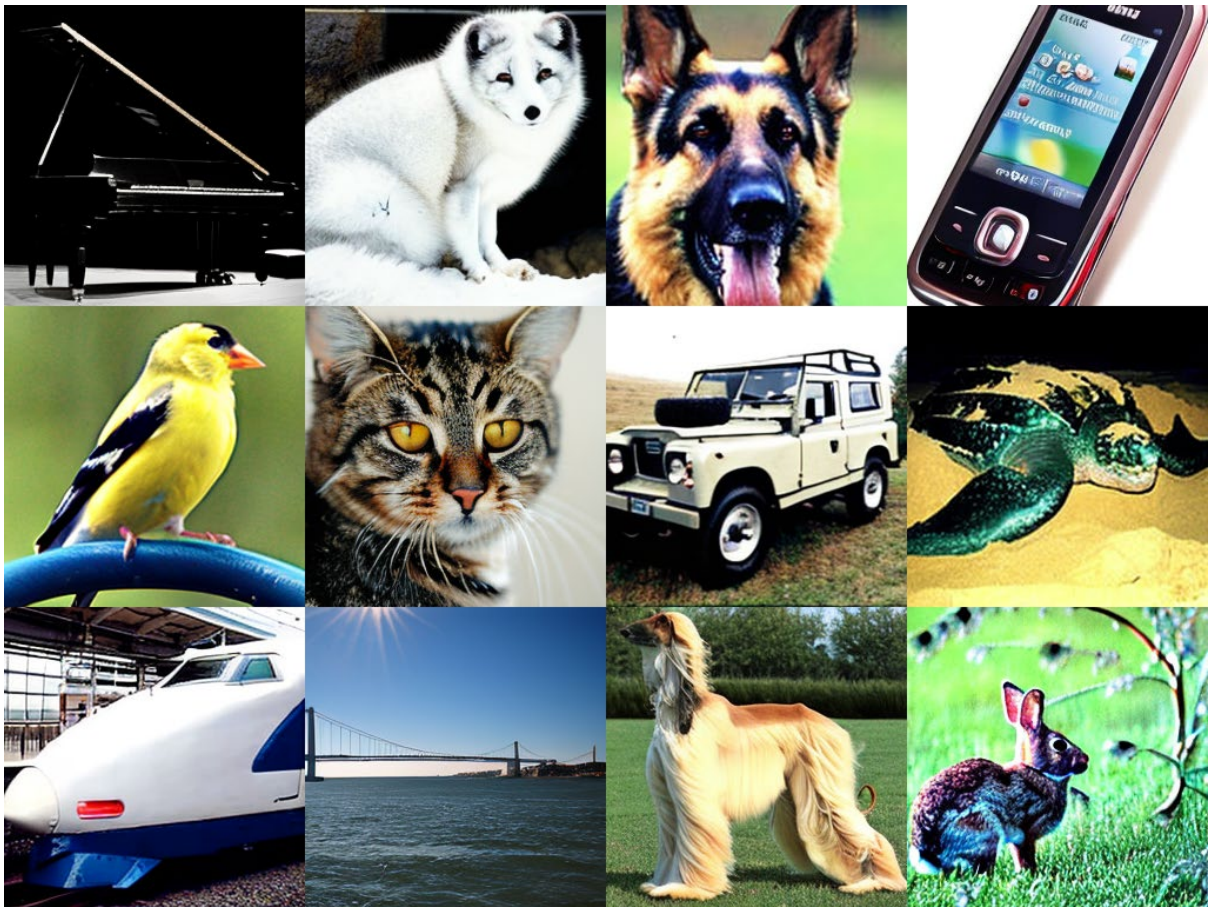

Figure 3. Results generated by some models, the image size is 256 × 256
图 3. 部分模型所生成的结果展示，图像尺寸均为 256 × 256

图 3 展示了我们的生成模型在 ImageNet 256 × 256 数据集上训练的部分推理结果。可以看到，推理结果图像有很高的生成质量，比如说钢琴、手机和桥梁的图像，物体的边缘非常锐利，同时物体的阴影有带有一定的模糊化，说明模型具备锐化主要物体变化的能力；一些动物的毛发、水面的波纹以及细碎的草地，表明模型具备复杂细节的生成能力，很好地学习到原有物体的特征。这些生成结果进一步验证了我们的模型具有很强的复杂图像的推理和生成能力。

4.2. 分块生成策略研究

在本次实验中，我们对图像生成模型使用分块生成的策略，通过计算不同分块的空间复杂度来自适应调整生成模型的层数，以达到优化生成图像质量和加快模型推理速度的目的。

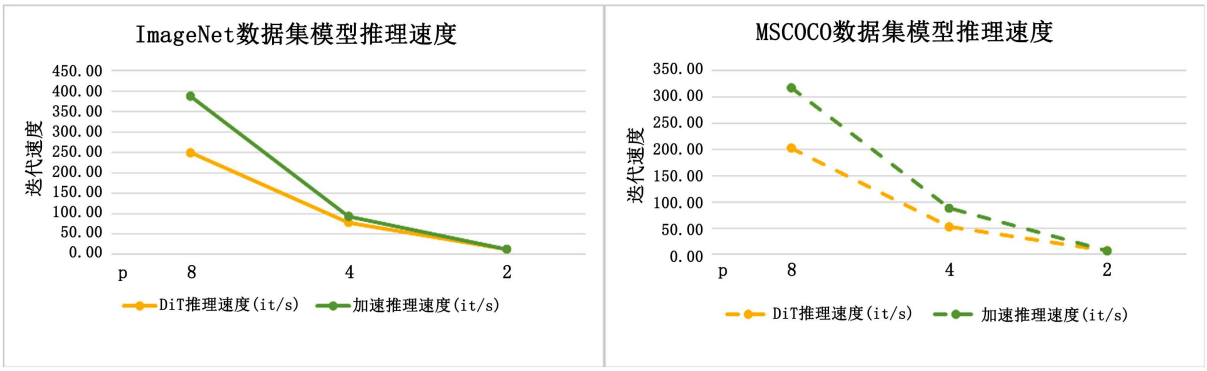

Figure 4. Comparison of image inference iteration speeds of the DiT architecture model and the model using a complex image acceleration inference strategy under the same test environment
图 4. 同一测试环境下 DiT 架构模型和采用复杂图像加速推理策略的模型在图像推理的迭代速度对比图

如图 4 所示，我们在两个数据集上分别用 DiT-L 模型和采用分块生成策略的图像生成模型进行对比，两个模型在相同的实验设备和计算环境下，测定推理同类别的一张图像的时间，生成图像的尺寸统一为 256 × 256。可以观察到，在分块较多的时候模型可以有效减少一张图像的推理时间，原因在于经过比较细致的分割的子图像多数空间复杂度较低，根据分块生成策略该部分的模型深度和潜空间(LDM)的尺寸也较小，而空间复杂度较高的部分和原始模型的结构大致相当，总体上来说，推理一张图像所需要的计算资源减少，推理速度加快。分块较少的情况下，子图像所承载的信息量较多，推理速度的提升并不明显。

4.3. 未来研究

本文围绕图像空间复杂度计算和自适应推理模型的深度生成策略展开，探讨了复杂图像加速计算生成的可能性。在空间复杂度的计算上我们仅考虑了一张图像和其高斯模糊后的图像相对比而量化产生的图像复杂度结果，而忽视了图像生成模型的一部分条件输入——提示词或者是类别标签，比如说生成一支笔的图像复杂度一定不及生成人物面部的图像复杂度，我们将在之后的研究中进一步探索这样的可能性。

此外，对于复杂度档次的划分和推理模型深度的设定，我们过多地依赖人为的设定，没有有效地探索更多的设定可能性对于实验结果的影响，这些内容我们将在之后的研究中跟进。

5. 总结

本文针对文本到图像生成领域中普遍存在的推理效率低下及复杂场景生成质量不足的问题，提出了

一种基于图像空间复杂度的自适应生成策略。该方法通过综合纹理与结构特征计算图像复杂度，并将图像分块处理，动态调整生成模型的推理深度，实现了在保持图像质量的同时，显著提升生成速度。实验结果表明，在 ImageNet 256 × 256 和 MSCOCO 2017 数据集上，所提方法在 FID、sFID 及 IS 指标上均优于现有主流模型，生成图像的主要物体边缘清晰、细节丰富，验证了该方法在复杂图像生成任务中的有效性。进一步分析了分块策略、复杂度划分及深度调整对生成性能的影响，为自适应生成模型的优化提供了实践依据。

参考文献

[1] Yin, M. and Li, J. (2023) A Systematic Review on Digital Human Models in Assembly Process Planning. *The International Journal of Advanced Manufacturing Technology*, **125**, 1037-1059. https://doi.org/10.1007/s00170-023-10804-8

[2] Brack, M., Friedrich, F., Kornmeier, K., Tsaban, L., Schramowski, P., Kersting, K., *et al.* (2024) LEDITS++: Limitless Image Editing Using Text-to-Image Models. 2024 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 16-22 June 2024, 8861-8870. https://doi.org/10.1109/cvpr52733.2024.00846

[3] Liu, V., Vermeulen, J., Fitzmaurice, G. and Matejka, J. (2023) 3DALL-E: Integrating Text-to-Image AI in 3D Design Workflows. Proceedings of the 2023 *ACM Designing Interactive Systems Conference*, Pittsburgh, 10-14 July 2023, 1955-1977. https://doi.org/10.1145/3563657.3596098

[4] Reed, S.E., Akata, Z., Yan, X., Logeswaran, L., Schiele, B. and Lee, H. (2016) Generative Adversarial Text to Image Synthesis. *International Conference on Machine Learning*, New York, 19-24 June 2016, 1060-1069.

[5] Ye, S., Wang, H., Tan, M. and Liu, F. (2023) Recurrent Affine Transformation for Text-to-Image Synthesis. *IEEE Transactions on Multimedia*, **26**, 462-473.

[6] Sauer, A., Karras, T., Laine, S., Geiger, A. and Aila, T. (2023) Stylegan-T: Unlocking the Power of Gans for Fast Large-scale Text-to-Image Synthesis. *International Conference on Machine Learning*, Honolulu, HI, 12-15 July 2023, 30105-30118.

[7] Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B. (2022) High-Resolution Image Synthesis with Latent Diffusion Models. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 10684-10695. https://doi.org/10.1109/cvpr52688.2022.01042

[8] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. and Chen, M. (2022) Hierarchical Text Conditional Image Generation with Clip Latents. arXiv: 2204.06125.

[9] Ho, J., Jain, A. and Abbeel, P. (2020) Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems*, **33**, 6840-6851.

[10] Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C. and Zhu, J. (2022) DPM-Solver: A Fast Ode Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps. arXiv: 2206.00927.

[11] Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C. and Zhu, J. (2022) DPM-Solver++: Fast Solver for Guided Sampling of Diffusion Probabilistic Models. arXiv: 2211.01095.

[12] Salimans, T. and Ho, J. (2022) Progressive Distillation for Fast Sampling of Diffusion Models. arXiv: 2202.00512.

[13] Meng, C., Rombach, R., Gao, R., Kingma, D., Ermon, S., Ho, J., *et al.* (2023) On Distillation of Guided Diffusion Models. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 14297-14306. https://doi.org/10.1109/cvpr52729.2023.01374

[14] Song, Y., Dhariwal, P., Chen, M. and Sutskever, I. (2023) Consistency Models. arXiv: 2303.01469.

[15] Wang, Z., Xia, X., Chen, R., Yu, D., Wang, C., Gong, M., *et al.* (2025) LaVin-DiT: Large Vision Diffusion Transformer. 2025 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 10-17 June 2025, 20060-20070. https://doi.org/10.1109/cvpr52734.2025.01868

[16] Jia, W., Huang, M., Chen, N., Zhang, L. and Mao, Z. (2025) D^2iT: Dynamic Diffusion Transformer for Accurate Image Generation. 2025 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 10-17 June 2025, 12860-12870. https://doi.org/10.1109/cvpr52734.2025.01200

[17] Fang, G., Li, K., Ma, X. and Wang, X. (2025) Tinyfusion: Diffusion Transformers Learned Shallow. 2025 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 10-17 June 2025, 18144-18154. https://doi.org/10.1109/cvpr52734.2025.01691

[18] Wang, Z. and Bovik, A.C. (2009) Mean Squared Error: Love It or Leave It? A New Look at Signal Fidelity Measures. *IEEE Signal Processing Magazine*, **26**, 98-117. https://doi.org/10.1109/msp.2008.930649

[19] Girod, B. (1993) What's Wrong with Mean-Squared Error. In: Watson, A.B., Ed., *Digital Images and Human Vision*, The

MIT Press, 207-220.

[20] Wang, Z., Bovik, A.C., Sheikh, H.R. and Simoncelli, E.P. (2004) Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, **13**, 600-612. https://doi.org/10.1109/tip.2003.819861

[21] Sheikh, H.R. and Bovik, A.C. (2006) Image Information and Visual Quality. *IEEE Transactions on Image Processing*, **15**, 430-444. https://doi.org/10.1109/tip.2005.859378

[22] Chandler, D.M. (2010) Most Apparent Distortion: Full-Reference Image Quality Assessment and the Role of Strategy. *Journal of Electronic Imaging*, **19**, Article ID: 011006. https://doi.org/10.1117/1.3267105

[23] Prashnani, E., Cai, H., Mostofi, Y. and Sen, P. (2018) PieAPP: Perceptual Image-Error Assessment through Pairwise Preference. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 1808-1817. https://doi.org/10.1109/cvpr.2018.00194

[24] Popat, K. and Picard, R.W. (1997) Cluster-Based Probability Model and Its Application to Image and Texture Processing. *IEEE Transactions on Image Processing*, **6**, 268-284. https://doi.org/10.1109/83.551697

[25] Balle, J., Stojanovic, A. and Ohm, J. (2011) Models for Static and Dynamic Texture Synthesis in Image and Video Compression. *IEEE Journal of Selected Topics in Signal Processing*, **5**, 1353-1365. https://doi.org/10.1109/jstsp.2011.2166246

[26] Peebles, W. and Xie, S. (2023) Scalable Diffusion Models with Transformers. 2023 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Paris, 1-6 October 2023, 4172-4182. https://doi.org/10.1109/iccv51070.2023.00387

[27] Perez, E., Strub, F., De Vries, H., Dumoulin, V. and Courville, A. (2018) Film: Visual Reasoning with a General Conditioning Layer. *Proceedings of the AAAI Conference on Artificial Intelligence*, **32**, 3942-3951. https://doi.org/10.1609/aaai.v32i1.11671

[28] Brock, A., Donahue, J. and Simonyan, K. (2019) Large Scale GAN Training for High Fidelity Natural Image Synthesis. arXiv: 1809.11096.

[29] Karras, T., Laine, S. and Aila, T. (2019) A Style-Based Generator Architecture for Generative Adversarial Networks. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 4396-4405. https://doi.org/10.1109/cvpr.2019.00453

[30] Dhariwal, P. and Nichol, A. (2021) Diffusion Models Beat GANs on Image Synthesis. 35*th Conference on Neural Information Processing Systems* (*NeurIPS* 2021), Sydney, 6-14 December 2021, 8780-8794.

[31] Zhang, R., Isola, P., Efros, A.A., Shechtman, E. and Wang, O. (2018) The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 586-595. https://doi.org/10.1109/cvpr.2018.00068

[32] Ding, K., *et al.* (2020) Image Quality Assessment: Unifying Structure and Texture Similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **44**, 2567-2581.

[33] Deng, J., Dong, W., Socher, R., Li, L., Kai Li, and Li Fei-Fei, (2009) ImageNet: A Large-Scale Hierarchical Image Database. 2009 *IEEE Conference on Computer Vision and Pattern Recognition*, Miami, 20-25 June 2009, 248-255. https://doi.org/10.1109/cvpr.2009.5206848

[34] Lin, T.Y., Maire, M., Belongie, S., *et al.* (2014) Microsoft COCO: Common Objects in Context. In: Fleet, D., Pajdla, T., Schiele, B. and Tuytelaars, T., Eds., *Lecture Notes in Computer Science*, Springer, 740-755.

[35] Loshchilov, I. and Hutter, F. (2017) Decoupled Weight Decay Regularization. arXiv: 1711.05101.