arXiv:2304.01515v1 [cs.LG] 4 Apr 2023

用于掩码生成模型文本到图像生成的文本条件采样框架

Jaewoong Lee*
KAIST
大韩民国
hello3196@kaist.ac.kr

Sangwon Jang*
延世大学
大韩民国
agwmon@gmail.com

Jaehyeong Jo、Jaehong Yoon
KAIST
大韩民国
{harryjo97, jaehong.yoon}@kaist.ac.kr

Yunji Kim、Jin-Hwa Kim、Jung-Woo Ha
NAVER AI实验室
大韩民国
{yunji.kim, jinhwa.kim, jungwoo.ha}@navercorp.com

Sung Ju Hwang
KAIST
大韩民国
sjhwang82@kaist.ac.kr

摘要

基于词元的掩码生成模型因并行解码带来的快速推理时间而日益受到青睐。尽管近期的基于词元的方法已取得与基于扩散的模型相当的性能，但由于它们在同时采样多个词元时没有考虑词元之间的依赖关系，生成性能仍不理想。我们通过实证研究探讨了这一问题，并提出一种可学习采样模型——文本条件词元选择（*TCTS*），利用文本信息进行局部监督，以选择最优词元。*TCTS*不仅提升了图像质量，也改善了生成图像与给定文本之间的语义对齐。为进一步提升图像质量，我们提出一种连贯的采样策略——频率自适应采样（*FAS*），并将其应用于根据自注意力图划分的各个词元组。我们在多种生成任务中验证了*TCTS*与*FAS*结合的有效性，结果表明，该方法在图文对齐和图像质量方面显著优于基线方法。我们的文本条件采样框架无需修改原有生成模型，即可将推理时间进一步缩短*50*% 以上。

1. 引言

在视觉领域生成式人工智能系统蓬勃发展的浪潮中，文本条件图像生成 [25, 32, 41] 近年来正逐渐崭露头角。尽管许多近期研究已成功合成出高质量图像 [18, 33] ，并在类别条件场景中实现了合理的类别对齐 [4, 8], ，但文本到图像（T2I）生成更具挑战性，因为生成在语义上与输入文本对齐的视觉结果并非易事。我们大致可以将文本到图像生成的相关研究分为基于 Transformer 的自回归（AR）[11, 29, 32] 和基于扩散 [17, 28] 两类方法。随着语言模型的发展，采用 Transformer 的 AR 模型在文本到图像生成方面展现出令人瞩目的性能。尽管取得了成功，它们仍存在单向偏差这一不利于图像生成的问题；更关键的是，与现有模型相比，其采样过程所需时间超过 10 倍。另一类研究是基于扩散的方法，旨在通过对噪声样本进行迭代去噪来生成图像。尤其是，若干连续扩散模型[10, 31] 已取得出色的性能并降低了计算成本。然而，在推理过程中，为了获得高质量图像，它们仍需要过多的采样步骤。

最近，一类被称为基于令牌的扩散模型 [6, 22, 40], 的新型生成模型应运而生，旨在解决文本到图像生成问题。基于令牌的扩散模型将潜在特征量化为词元，并在离散状态空间中应用分类损坏过程 [1],，而传统扩散模型则在连续空间中使用高斯噪声。在各种离散扩散方法中，最常使用的是基于掩码的扩散，这与 [1], 中使用的吸收态扩散类似。与现有的 AR 模型相比，这种基于令牌的方法能够通过同时采样多个词元来加速生成过程，因此具有优势。尽管重建能力受到限制，这些基于令牌的扩散模型在 FID 分数方面仍显著优于竞争方法，而且所需采样步骤少于连续扩散模型 [5, 36]。

*贡献相同。初步工作。



图1：**MS-COCO 数据集上的生成样本及各种采样方法的评估图，展示了它们之间的权衡**。与其他方法相比，均匀采样是一种固定策略，文本对齐效果明显较差（FID-40K：15.61，MID-L：21.23）。随机撤销采样是一种可撤销策略，能够改善文本对齐效果（FID-40K：16.81，MID-L：26.98）。我们的方法是将 TCTS 与 FAS 结合；与基线方法相比，图像质量和文本对齐效果均显著更好（FID-40K：13.6，MID-L：29.5）。指标基于全部 40K 张图像及其对应的单条图注计算。所有采样方法的无分类器引导尺度均固定为 5。

然而，一次采样多个词元通常会导致生成图像前后不一致 [36]。对于每个位置，生成器都会输出彼此协调的概率分布。然而，无法保证从这些分布中采样出的每个词元都能完美地相互匹配。换言之，无论生成器的能力如何，它仍有可能采样出彼此不兼容的词元，从而产生潜在的无意义输出 [23, 36]。这导致采样步骤与生成质量之间存在权衡。减少采样步骤可以加快生成速度，但由于同时采样的词元数量较多，会导致性能下降。尤其是在文本到图像生成任务中，这一问题更加突出，因为与无条件图像或类别条件图像的分布相比，与文本对齐的图像分布更加受限。

为了解决这些局限性，我们提出了一种新颖的采样方法，根据文本条件在扩散过程中对图像进行细化，并将其称为文本条件词元选择 (TCTS)。TCTS 是一种采样策略，可以屏蔽并重新采样先前采样的词元，我们将其称为可撤销。为了找出与给定文本条件不对齐的词元，我们提出训练一个模型来选择需要屏蔽并重新生成的词元，使其与给定文本良好对齐。将这一方法与可撤销采样方案相结合后，与朴素生成模型相比，TCTS 能以更少的采样步骤生成高质量且文本对齐度更高的图像，如图 1所示。我们还提出了频率自适应采样 (FAS)，以解决在相对较长的扩散步数中应用可撤销方法时出现的过度简化问题。FAS 利用图像的自注意力图，并采用一种混合采样方法，我们称之为持久采样，以防止

这一问题。我们的贡献总结如下：

- 我们通过实验表明，可撤销采样策略对于文本对齐与图像质量之间的权衡至关重要，并且相较于以往的固定采样方法，我们提供了深入分析。

- 我们提出了一种基于可学习词元选择模型的新型可撤销采样方法，即使采样步数更少，也能显著提升文本对齐度和图像质量，而且无需重新训练生成器模型。

- 此外，我们提出了一种基于自注意力图的新型采样方法，可与 TCTS 结合使用，以解决过度简化问题。

2. 相关工作

已有多项研究致力于文本到图像生成任务，其中大多数基于生成对抗网络（GAN）[37, 39, 42, 43, 44, 45, 46]。然而，直接在图像像素空间中进行计算颇具挑战，而且由于涉及的像素数量庞大，生成高分辨率图像需要大量计算。因此，通常采用两阶段方法：模型首先将图像转换为一系列编码，然后预测这些词元化编码。

自回归模型继 Transformer [38] 和 GPT [27]，之后，自回归（AR）模型 [9, 29, 30, 41] 提出将图像词元化，并将文本到图像生成问题表述为将文本描述转换为图像词元的问题，这与将一种语言机器翻译成另一种语言颇为相似。随着图像词元化和语言模型的发展，

这种方法已取得很高的性能。然而，即使有分词阶段的帮助，仍有大量词元需要预测；由于自回归特性，生成时间因此显著延长。此外，错误会在不可撤销的过程中不断累积，并带来不理想的单向偏差。

潜空间扩散模型 近年来，扩散模型 [17,35] 为文本到图像生成引入了新的范式 [32]。它们通过逐步对噪声进行去噪来生成高质量图像，但需要迭代步骤和大量计算。两阶段方法 [10, 31] 提出在潜空间中执行扩散，在保持性能的同时成功降低计算量。然而，大多数方法仍需要约 50 个推理步骤才能展现出合理的性能；此外，由于在推理过程中会对整个图像进行全局去噪，因此局部修正图像需要复杂的处理流程 [31]。

基于令牌的扩散模型 基于令牌的扩散模型 [5, 12, 22] 在从文本生成复杂场景方面表现出色。不同于传统扩散模型将高斯噪声应用于连续空间，基于令牌的扩散模型会将潜在特征量化为词元，并在离散状态空间中应用类别式破坏过程 [1]。在各种离散扩散方法中，基于掩码的扩散因其吸引人的特性而得到最广泛的应用；它能够同时采样多个词元，因此与潜空间扩散模型相比，所需采样步骤更少、生成时间更短。然而，一次选择多个词元会导致整幅图像中的不一致，这通常被称为联合分布问题。此外，词元会受到已经修改的空间相邻词元的强烈影响，因此一个错位令牌可能会使错误扩散至整幅图像。为解决这一问题，一些工作 [6, 22] 修改了采样策略，而另一些工作 [23] 引入了新的模型来寻找错位令牌。

具体而言，Lezama等人 [23] 提出了一种可学习模型，可在每个步骤中移除错位令牌。尽管大多数现有研究会随机选择要丢弃的词元，但这项工作会检查所有采样得到的词元。它会输出每个位置的分数，该分数表示是否对应的词元服从真实分布。经过生成器的预测和采样过程后，模型根据对应词元的分数选择位置。

3. 文本条件令牌选择

掩码图像生成模型的扩散过程从一幅完全被掩码的图像开始，并为整幅图像采样 x_t 的词元，以预测 $\hat{x}_0$，同时

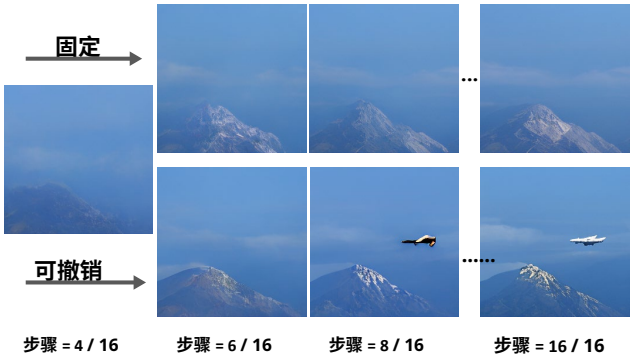


图2：**使用固定方法和可撤销方法进行扩散时的重建图像**。只有可撤销方法能够根据文本“天空中山脉上方的飞机尾部视角。”编辑词元以生成一架飞机。

将其中一部分重新掩码，以获得去噪样本x_{t-1}。现有的掩码图像生成模型使用均匀采样或基于置信度的采样策略来确定重新掩码的位置：均匀采样 [12, 22] 会从所有位置的预测词元中随机选择要保留词元的位置，而基于置信度的策略（例如纯度采样 [36],）则根据置信度分数选择位置，置信度分数定义如下：

$$\mathrm{CS}(i,t) = \max_{j=1,\cdots,K} p(x_0^i = j | x_t^i), \tag{1}$$

其中，t 是步骤编号，K 是代码本的大小，i 是词元的位置。直观而言，置信度分数越高，意味着生成器越确信应在对应位置采样该词元。通过保留这些置信度高的词元并丢弃含糊不清的词元，纯度采样在实践中表现良好 [5]。

然而，这些策略不可撤销，即，固定，这意味着词元一旦被采样，此后便无法修改，即使它们与给定文本不匹配或与其他词元不对齐也是如此。一旦错位令牌在过程中被固定，错误就会不断累积并扩散到整幅图像中。我们观察到，固定策略会对文本对齐产生负面影响，如图1所示。

3.1. 随机撤销采样策略

为了解决这一问题，我们首先提出一种名为随机撤销采样的新采样策略，并将其作为最终方法的基线。它与均匀采样类似，二者都通过从均匀分布中采样来确定位置。不过，均匀采样是一种固定策略，而随机撤销采样会从整个图像空间中选择位置，不受先前固定位置的限制，因此整个过程具有可撤销性。它会随机选择一定数量的词元，使其保留到下一步骤，并撤销先前固定的词元。可撤销这一特性有助于使图像更好地与给定的文本信息对齐，尤其是在复杂场景中。

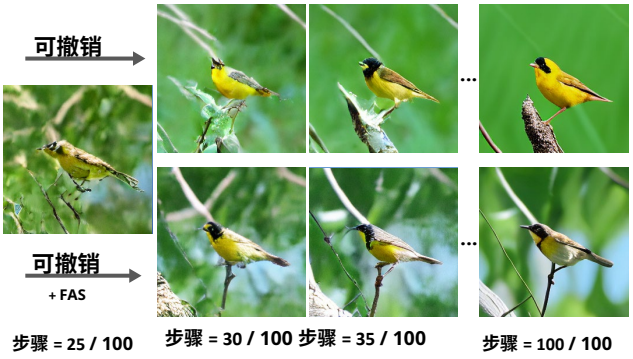


图3：**使用带有和不带 FAS 的可撤销方法进行扩散时的重建图像**。如底部行所示，顶部行图像的背景过度简化，而我们提出的 FAS 可防止这一问题。文本为“这只小鸟整体呈灰色，背部和胸部带有黄色斑点，腹部还有些许白色。”。

具体而言，如图2上排所示，固定策略在纠正错误或生成多样化图像方面存在局限，因为它们无法重新生成错误的词元，也无法修改缺失的重要对象或属性。相比之下，可撤销策略能够提供纠正无效生成词元的机会，从而增强文本对齐效果，如图2下排所示。由于可撤销策略不受移除或修正哪些词元的限制，因此有助于缓解误差累积。

然而，随机撤销采样存在两个主要问题，需要加以解决。第一个问题是其不稳定性：先前采样的词元会被随机丢弃，而这些词元曾作为该步骤的给定条件。这可能会降低输出质量，如图1所示。

第二个问题，即过度简化，是在涉及较长扩散步数的生成过程中使用该方法时产生的。与生成更逼真背景的固定方法相比，输出图像中的低频部分（例如背景）往往会过度简化。在图 3中可以看到过度简化的样本，这也会显著影响 表 1 和 2中的 FID指标。我们观察到，这是由于在大量步骤中反复进行重新采样，而固定调度不会对任何词元进行重新采样。更多的重新采样机会会使词元趋向于更安全、更简单的模式。如图 3顶部行所示，即使经过多次重新采样过程，具有丰富视觉细节的高频区域也不会过度简化；但由于过度重新采样，低频区域会迅速变得缺乏多样性，且其模式更加简单。

为了确认文本对齐与图像质量之间的这种权衡，我们将持久采样定义为均匀采样与随机撤销采样之间的一种插值。具体而言，持久权重会提高保留——

算法 1持久采样

输入： k_t：每个步骤采样的词元数量 t， w：持久权重， N：所有词元的数量

$x_T \leftarrow [[\text{掩码}]]_N$
对于 $t = T, T-1, \ldots, 1$ **执行**
  $\hat{x}_0 = G_\theta(x_t, c)$
  $A_t = \{i | x_t^i = [\text{掩码}]\}$
  $m = k_T + k_{T-1} + \ldots + k_t$
  $\mathcal{U}_t(I = i | i \in A_t) : \mathcal{U}_t(I = i | i \in A_t^C) = 1: w$
  $\rightarrow$均匀分布 + 持久权重
  $i_1, i_2, \ldots\ldots i_m \sim \mathcal{U}(I)$ $\leftarrow$无放回抽样
  $x_t \leftarrow [[\text{掩码}]]_N$
  对于 $i = i_1, i_2, \ldots\ldots i_m$ **执行**
    $x_t^i \leftarrow \hat{x}_0^i$
  结束循环
结束循环
返回： 生成图像 x_0

上一步中已采样词元的概率。随机撤销采样为所有位置分配相同的概率，而持久采样则将持久权重乘以先前保留位置的概率。更多细节见算法1。如果持久权重 w 为 1，那么所有词元被固定的概率都相同，例如随机撤销采样。如果 w 足够大，那么先前采样的词元会一直保留到最后，例如均匀采样。在图1中可以看到这种权衡：采样策略越接近随机撤销采样，文本对齐效果越好，而图像质量越差。我们还在第 4.5节进一步分析这些采样方法的影响。

3.2. 文本条件词元选择

我们提出了一种名为 文本条件令牌选择 （TCTS）的可学习模型。TCTS 经过训练后，能够输出得分图，以检测在给定文本条件下不符合真实分布的词元，并选择图像中与文本或其他词元未对齐的词元。通过在每个步骤屏蔽所选词元，扩散网络会接收高质量、与文本对齐的词元作为输入，从而采样下一组词元，缓解以往采样策略中的误差累积和联合分布问题。此外，由于 TCTS 经过训练能够区分对齐良好的词元，因此不同于其他可撤销方法，它避开了前文提到的不稳定性问题。从 第 4 节的实验结果可以看出，TCTS 提升了文本对齐效果，同时不会损害图像质量。

频率自适应采样（FAS）解决较长扩散步数下过度简化问题的最简单方法，是采用持久采样来减少对词元的过度重新采样。然而，正如前文所述，这种

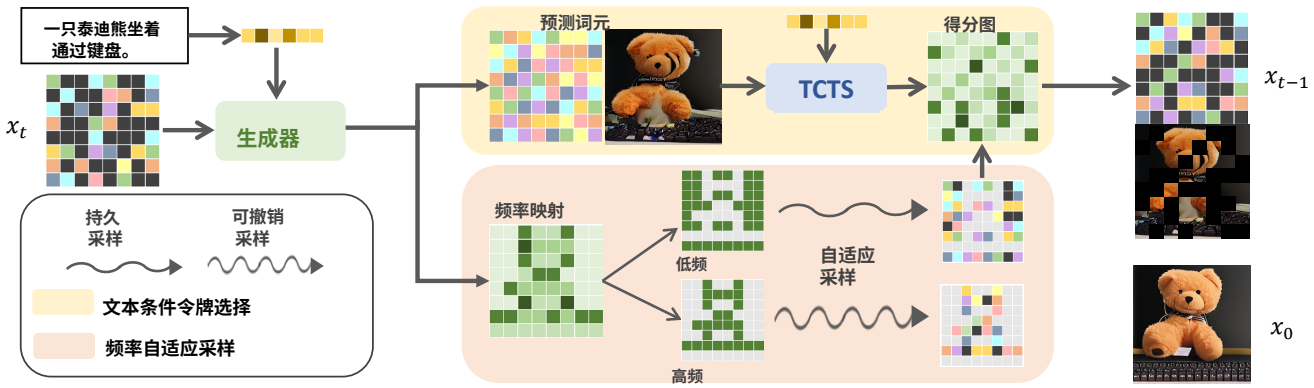


图4：**所提出的 TCTS 和 FAS 的整体生成框架**。生成器预测词元后，TCTS 利用文本条件检测错位令牌，并输出得分图。同时，FAS 根据生成器的自注意力图按频率划分词元，对高频划分执行可撤销采样，对低频划分执行持久采样。自适应采样预测 $\hat{x}_0$，并根据 x_t 决定要掩码的若干位置。FAS 生成的令牌图与得分图相结合，以预测 x_{t-1}。经过迭代过程后，我们的模型移除所有掩码并生成 x_0。

Step	方法	MID-L ↑	SOA-I ↑	CLIP-S ↑	FID-30K ↓
16	纯度	11.02	72.38	0.2474	19.20
	Uniform	17.94	74.80	0.2500	16.17
	RR	23.60	78.79	0.2526	17.10
	TCTS +FAS	**26.72**	**79.52**	**0.2559**	**14.45**
25	纯度	16.84	75.21	0.2487	18.39
	Uniform	22.27	77.08	0.2524	15.91
	RR	26.77	**81.10**	0.2543	18.43
	TCTS +FAS	**27.79**	80.87	**0.2563**	**15.39**

表1：**在 MS- COCO 数据集上的采样方法定量评估**。MS-COCO 图像-图注对的真实 MID 为 54。73。所有采样方法的无分类器引导尺度均固定为 5。

Step	方法	MID-L ↑	CLIP-S ↑	FID ↓
16	纯度	-24.21	0.2410	15.21
	Uniform	-25.60	0.2404	16.57
	RR	-25.03	0.2371	17.38
	TCTS + FAS	**-19.88**	**0.246**	**12.35**
25	纯度	-21.26	0.2384	**12.60**
	Uniform	-23.04	0.2396	13.02
	RR	-23.29	0.2364	14.53
	TCTS + FAS	**-18.31**	**0.2409**	13.67

表2：**在 CUB数据集上的采样方法定量评估**。CUB图像-标题对的真实MID为 15.85。由于 CUB数据集中的每张照片只有一只鸟，因此该指标的意义有所降低，本表省略了 SOA。

方法可能会轻微牺牲文本对齐，并限制通过重新采样来修正图像的机会。为进一步改善这种权衡，我们提出了一种名为频率自适应采样（*FAS*）的新方法，可应用于 TCTS。

FAS 是一种利用生成器的自注意力图，将重新采样限制在图像低频区域的方法。在 [20], 已知生成器的自注意力层包含图像的频率信息，这一现象在基于令牌的扩散模型中也有所观察。我们利用这一特性，无需额外操作即可区分图像的频率区域，并使用持久权重仅在低频区域进行持久采样。如图 3底部一行所示，可撤销采样与 FAS 方法相结合，使需要丰富视觉细节的区域能够反复重新采样，同时限制相对简单区域的重新采样次数，从而避免过度简化。详细算法见补充材料A。我们模型的整体框架（包括 TCTS 和 FAS）如图4所示。

4. 实验

为验证我们的令牌选择框架的有效性，我们对 [36] 中的 Transformer 进行了修改，使其在 MS-COCO 数据集上包含 3.1 亿个参数，在 CUB 数据集上包含 8300 万个参数。我们从 CLIP ViT-B/32 [26] 中提取文本特征，用于文本条件生成。在 TCTS 的训练过程中，我们冻结生成器的参数，并独立训练我们的模型，使其适用于其他类型的基于令牌的扩散模型。我们使用二元交叉熵损失作为目标函数，并在 MS-COCO [24] 和 CUB [13]数据集上对模型训练 200 个周期，以验证我们的方法和基线方法。我们通过对引导强度进行随机采样来利用无分类器引导 [19]，以提升生成质量。我们还在补充材料A中进一步介绍架构、超参数，以及关于使用无分类器引导的补充分析。

模型	MID-L ↑	MID-B ↑	SOA-I ↑	FID-30K ↓
GLIDE [25]	1.03	1.00	-	32.08
AttnGAN [39]	-65.20	-8.90	39.01	29.15
DM-GAN [22]	-44.66	3.51	48.03	22.90
DF-GAN [37]	-58.75	-15.21	-	31.75
VQ-Diffusion [40]	-19.63	5.77	-	13.13
LAFITE [45]	6.26	35.17	74.78	**8.03**
TCTS +FAS	**26.02**	**38.98**	**79.52**	9.75

表3：**所提出方法与近期文本到图像生成模型在 MS-COCO 数据集上的比较**。MID-L 和 MID-B 分别使用 ViT-L/14 和 ViT-B/32 计算。MS-COCO 图像-标题对的真实MID-L 和真实 MID-B 分别为 54.73 和 41.57。我们在 16 步设置下评估所提出的方法，并采用 3 到 5 的无分类器引导。

指标 由于 FID [15] 指标被认为存在问题，因为它无法准确表示图像质量， [5]也无法考虑文本对齐，因此我们还使用互信息散度（MID） [21] 评估生成图像的文本对齐。该指标支持逐样本评估，且在人类李克特量表判断相关性方面优于以往指标。具体而言，由于 MID 对使用误导性图像描述生成的图像反应更加敏感，[34], 因此它更适合分析复杂且具有挑战性的图注中的文本对齐。此外，我们还使用了 CLIP 分数 [14] 和 SOA-I [16], ，它们广泛应用于文本到图像合成。请注意，所有实验中的 CLIP 分数均使用 ViT-L/14 计算。

4.1. 文本到图像合成

图1 左侧和中间展示了使用改进的 VQ 扩散 [36]和不同采样方法生成的图像。均匀令牌采样生成的示例与给定文本的对齐效果较差，并且通常会出现物体划分混乱的情况，因为在同一步骤选出的预测词元无法适当地反映其他词元的变化（例如，破碎的甜甜圈，以及不完整的椅子和海鸥）。随机撤销（RR）采样可以重新更新此前采样的词元，通过编辑掉不合逻辑的区域，并在每个生成步骤中迭代地考虑整个场景，似乎能够改善与文本的对齐效果。然而，RR 以完全随机的方式选择词元，不受任何约束，导致生成质量以及与文本的对齐效果不佳（例如，消失的男人和火车）。另一方面，我们的方法通过在采样阶段选择文本条件化的词元，缓解了 RR 所带来的问题，并成功生成了高质量图像，这些图像与给定文本图注之间具有清晰的语义联系。

这一现象可以通过表 1定量观察到。随机撤销采样的16步生成，比其他固定



图 5：**使用 TCTS 进行图像细化的结果**。上：原始样本，下：使用 TCTS 扩散 8 步后的精炼图像。

方法	MID-L ↑	FID ↓
原始	12.10	**16.44**
RR	13.64	17.18
TCTS	**15.64**	16.92
原始	4.54	19.13
随机掩码	6.06	19.12
TCTS 掩码	**6.47**	**18.64**

表4：**精炼图像的定性评估**。上：通过增加修订步骤进行精炼。下：通过掩码得分最低的词元进行精炼。

采样方法的25步生成甚至表现出更好的文本对齐。将各方法与其25步版本进行比较，还可以观察到 CLIP 分数与 FID 之间的权衡。我们的方法在大多数指标上都优于其他基线，尤其是将 MID-L 与纯度和 Uniform 等固定采样方法进行比较时。由于我们的方法能够在生成过程中进行修正，因此只需更少的扩散步数，就能达到基线采样方法的性能；进一步分析见第 4.5节。我们还在表 2中展示了参数量较少时在 CUB数据集上的实验结果，性能令人满意。不过，CUB数据集中的每张图像仅包含单个物体，因此表现出与前文所述略有不同的趋势。在表3中，我们将本方法与其他文本生成图像模型进行了比较。我们的方法在文本对齐指标上大幅优于其他模型，尤其是在 MID-L 指标上，并且其 MID-B 值接近真实图像。更多由我们的方法生成的样本见补充材料B。

4.2. 图像细化

掩码图像生成模型的优势在于能够实现快速局部精炼。受 [22] 和[23]的启发，我们采用两种不同的方法进行图像细化。首先，我们使用 TCTS 对通过均匀采样生成的图像应用额外的修订步骤。整体图像

通过增加额外的细化步骤，可以提升质量，如图 5所示。为了展示我们的 TCTS 的图像细化性能，我们还测量了采用随机撤销采样进行图像细化的性能。虽然随机撤销采样确实提升了样本质量，但额外的细化步骤会增加 FID 分数，这是由与增加生成步骤类似的效应所导致的。

此外，我们使用 TCTS 对得分最低的 60% 词元进行掩码，并通过均匀采样生成图像。我们发现，精炼图像的所有指标均优于原始图像。在表4中，我们将其与不使用 TCTS、随机对词元进行掩码后精炼得到的样本进行比较。受篇幅限制，我们在补充材料B中介绍这两种精炼方法的更多细节，并展示更多样本。

4.3. 无掩码对象编辑

由于掩码图像生成模型具备局部精炼能力，因此只需随机对图像的一部分进行掩码，并在新的文本条件下对其重新采样，即可实现无需手动掩码的图像编辑。然而，为了保持图像的整体结构，这种方法需要较低的掩码比例和较多的重采样步骤。这可能导致不必要的区域发生显著变化，或如前所述出现过度简化问题。此外，即使进行许多步骤，由于周围词元的影响显著强于新文本条件，使用小掩码编辑大型对象仍然更具挑战性。

受频率自适应采样中自注意力图作用方式的启发，我们利用与待更改对象对应词语的交叉注意力图，为重新采样词元赋予权重，从而能够对相应位置进行重新采样。这样便能以更少的扩散步数高效地进行图像编辑，也更容易编辑较大的对象。虽然这与 DiffEdit 相似， [7], 但得益于掩码生成模型的局部精炼能力，我们的方法以更直接的方式运行，无需额外计算。在图6中，我们仅通过 10 个扩散步数添加 25% 的噪声来编辑对象。使用我们的方法，可以用更少的编辑步骤更快地完成编辑，而且不会受到过度简化问题的影响。更多样本见补充材料B。

4.4. 高分辨率图像合成

使用编码器和解码器的连续扩散模型（例如掩码图像生成模型）可以生成——
生成训练集之外更大的图像。与 Bond-Taylor 等人类似 [3], ，我们根据模型输入尺寸将词元划分为若干子集，分别将其输入模型，然后对词元进行空间聚合，以合成高分辨率图像。通过这种方式，无需额外训练，我们就能使用相同的 TCTS 模型生成更逼真、更高分辨率的样本。此-

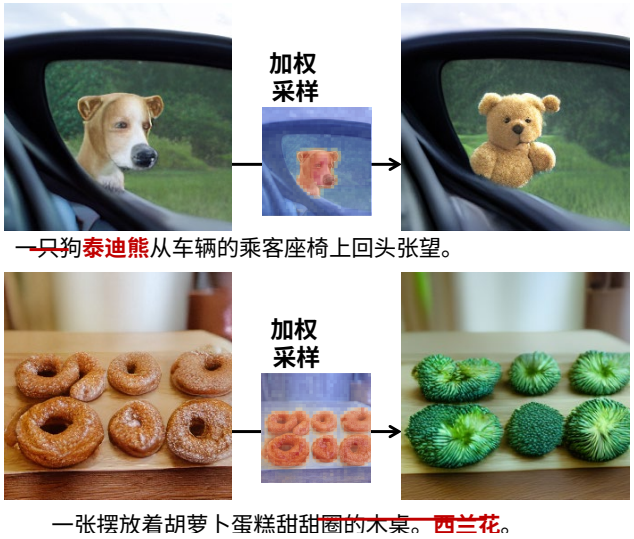


图 6：**无掩码编辑样本及交叉注意力图示例**。交叉注意力图与 TCTS 的得分图相乘，以进行加权采样。



图 7：**高分辨率（512^2）样本**。**左侧**：使用全部掩码标记生成的图像，**右侧**：使用 TCTS 作为超分辨率单元生成的图像。

外，我们提出了一种仅使用低分辨率 TCTS 生成高分辨率图像的新方法。具体而言，我们首先使用 TCTS 生成一张小尺寸图像，并以双三次模式对标记图进行上采样，使其达到所需尺寸。然后，我们将高分辨率标记图划分为重叠的小尺寸区块，并对所有区块进行多次细化。随后，我们便能使用低分辨率 TCTS 生成高分辨率图像，生成的样本展示于图7。

4.5. 进一步分析

早期采样策略的重要性如第3.1节所述，通过控制持久权重，我们可以在两种基线采样方法之间进行插值：均匀采样和随机撤销采样。我们设计了多种实验，以探究采样

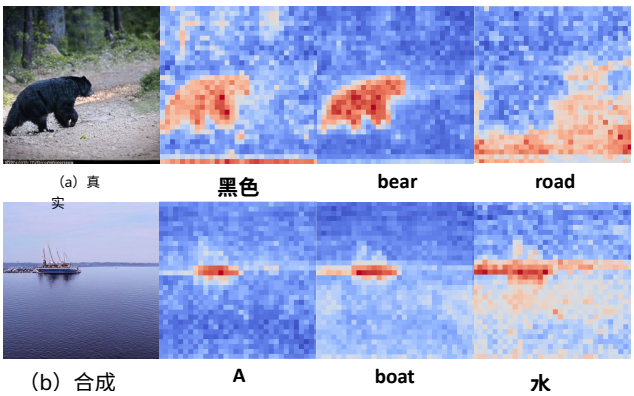


图 8：**TCTS 中每个词的平均交叉注意力图可视化。上方**：来自 COCO 数据集的图像：“一只大型黑熊正沿着一条肮脏的道路行走。”**下方**：合成图像：“一艘小船位于一片狭窄的水域中。”

方法对时间步的影响。我们在扩散步数期间切换采样方法，从均匀采样切换到随机撤销采样（U2R），反之亦然（R2U）。从图1最右侧的图中可以观察到，U2R的文本对齐性能优于R2U。

这些结果表明，在掩码比例较高的早期阶段进行采样，会显著影响最终图像的文本对齐；在连续扩散模型中也观察到了类似现象 [2]。先生成整幅图像，再通过若干修订步骤进行处理，例如[22]，或许有助于提升图像质量，但无法显著改善根本性的文本对齐。因此，要获得与文本良好对齐的图像，就必须在生成词元数量尚少的早期阶段谨慎采样。这一分析也体现在可撤销策略的结果中：该策略通过多次积极地重新采样词元，取得了更好的性能。具体而言，由于生成性能较差的弱生成器更容易受到联合分布问题的影响，在早期阶段给予其足够多的恢复机会，以丢弃词元并生成新的词元，是实现理想文本对齐的关键。

文本条件采样 实验表明，TCTS 在增强图像生成中的文本对齐方面发挥着至关重要的作用。尽管我们在训练过程中没有使用任何文本专用损失，但模型仍能有效处理文本条件并达到最佳性能。为了验证 TCTS 对图像中文本相关部分的注意力是否准确，我们将每个词的平均交叉注意力层可视化，如图8所示。结果表明，TCTS 能够关注真实图像和合成图像中的每个词。通过在生成阶段帮助选择相关词元，TCTS 无需额外复杂的文本相关损失即可提升模型性能。

方法	MID-L ↑	SOA-I ↑	CLIP-S ↑	FID-30K ↓
RR	**26.77**	**81.10**	**0.2543**	18.43
随机撤销采样 +FAS	26.32	80.78	0.2539	16.01
随机撤销采样 +持久权重	24.23	79.14	0.2534	**15.96**
TCTS	**27.82**	**80.93**	**0.2565**	16.69
TCTS + FAS	27.79	80.87	0.2563	15.39
TCTS + 持久权重	25.12	79.51	0.2556	**15.12**

表5：**在 MS-COCO 数据集上，消融研究了在可撤销方法中使用 FAS 和持久权重的效果**。所有模型均在 25步设置下进行评估。在 MS-COCO 图像-标题对上的真实MID-L 为54。73。

Step	模型	MID-L ↑	FID-30K ↓	Time
50	Uniform	24.04	16.84	×1
	RR	26.55	21.12	
16	TCTS + FAS	**26.72**	**14.45**	×**0.45**

表 6：**我们的方法与基线方法及模型在 MS-COCO 数据集上的 MID/FID 相对推理时间**。我们将模型的推理时间表示为其他基线方法推理时间的倍数。与其他基线方法相比，我们的模型速度更快，同时在 MID-L 和 FID-30K 上表现更好。更多分析见补充材料 C。

FAS 的消融研究 如第 3.2节所述，FAS 决定是否为每个位置应用持久权重。如果整幅图像被视为低频图像，FAS 会为每个位置应用该权重，这与持久采样相同。另一方面，如果整幅图像被视为高频图像，结合 FAS 后不会在任何位置应用该权重，因此不会改变采样策略。为了进一步分析 FAS 的影响，我们在两种可撤销方法上对其进行了评估：随机撤销采样和 TCTS。将 FAS 附加到一种基础采样方法上，可以视为两种方法的混合：基础可撤销采样方法和持久采样方法。 表 5 清楚地展示了 FAS 与持久权重之间的关系。尽管结合 FAS 的方法没有表现出最佳性能，但其性能始终接近各自两个对比方法中表现较好者。结合 FAS 的方法无需额外训练，便能在文本对齐和图像质量之间实现良好的平衡。

推理时间我们的 TCTS 和 FAS 会在每个生成步骤执行额外运算，因此与其他令牌选择基线方法相比，推理时间略有增加。然而，在文本对齐和图像质量方面，相对推理时间显著缩短。在表6中，仅用 16 个步骤，我们的模型在 MID-L 和 FID-30K 上的表现便优于基线采样方法进行 50 步图像生成的结果，这意味着我们的模型只需 $\times 0.45$ 时间即可生成更好的样本。

5. 结论

本文探讨了掩码扩散过程中哪些因素会影响输出图像，并造成图像质量与文本对齐之间的权衡。通过实证研究，我们发现可撤销采样能显著提升文本对齐效果，但会降低生成图像的质量。为解决这一问题，我们提出一种简单的词元采样策略，称为文本条件令牌选择，将可学习方法与可撤销方法相结合，在无需更新大规模预训练生成器的情况下，推动图像质量与文本对齐之间权衡的边界。我们发现，以持久和可撤销的方式进行协同采样，出乎意料地缓解了生成背景中的过度简化问题。我们提出的方法可用作各种生成任务的图像细化工具，能够以极少的步骤快速生成高质量图像。

参考文献

[1] Jacob Austin、Daniel D. Johnson、Jonathan Ho、Daniel Tar- low 和 Rianne van den Berg。离散状态空间中的结构化去噪扩散模型。载于《神经信息处理系统进展》，第 17981–17993 页。Curran Associates, Inc.，2021年。1、3 [2] Yogesh Balaji、Seungjun Nah、Xun Huang、Arash Vahdat、Ji- aming Song、Karsten Kreis、Miika Aittala、Timo Aila、Samuli Laine、Bryan Catanzaro、Tero Karras 和 Ming-Yu Liu。ediff- i：采用专家去噪器集成的文生图扩散模型。arXiv 预印本 arXiv:2211.01324，2022。8[3] Sam Bond-Taylor、Peter Hessey、Hiroshi Sasaki、Toby P Breckon 和 Chris G Willcocks。释放 Transformer 的潜能：利用离散吸收扩散进行并行词元预测，从而基于向量量化编码快速生成高分辨率图像。载于《计算机视觉—ECCV 2022：第 17 届欧洲会议，特拉维夫，以色列，2022 年 10 月 23–27 日，论文集》第 XXIII 部分，第 170–188 页。Springer，2022。7[4] Arantxa Casanova、Marlene Careil、Jakob Verbeek、Michal Drozdzal 和 Adriana Romero Soriano。实例条件生成对抗网络。《神经信息处理系统进展》，34：27517–27529，2021年。1[5] Huiwen Chang、Han Zhang、Jarred Barber、AJ Maschinot、Jose Lezama、Lu Jiang、Ming-Hsuan Yang、Kevin Murphy、William T Freeman、Michael Rubinstein 等。Muse：通过掩码生成式变换器实现文生图生成。arXiv 预印本 arXiv:2301.00704，2023。1、3、6、11[6] Huiwen Chang、Han Zhang、Lu Jiang、Ce Liu 和 William T Freeman。Maskgit：掩码生成式图像变换器。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 11315–11325 页，2022。1、3[7] Guillaume Couairon、Jakob Verbeek、Holger Schwenk 和 Matthieu Cord。DiffEdit：利用掩码引导进行基于扩散的语义图像编辑，2022。7

[8] Prafulla Dhariwal 和 Alexander Nichol。《扩散模型在图像合成方面击败生成对抗网络》。神经信息处理系统进展，34:8780–8794，2021年。1[9] Patrick Esser、Robin Rombach 和 Bjorn Ommer。《驯服 Transformer 以实现高分辨率图像合成》。载于 IEEE/CVF 计算机视觉与模式识别会议论文集，第 12873–12883 页，2021 年。2[10] Wan-Cyuan Fan、Yen-Chun Chen、DongDong Chen、Yu Cheng、Lu Yuan 和 Yu-Chiang Frank Wang。《Frido：用于复杂场景图像合成的特征金字塔扩散》。arXiv 预印本 arXiv:2208.13753，2022 年。1、3[11] Oran Gafni、Adam Polyak、Oron Ashual、Shelly Sheynin、Devi Parikh 和 Yaniv Taigman。《Make-a-scene：利用人类先验的基于场景的文本生成图像》。arXiv 预印本 arXiv:2203.13131，2022 年。1[12] Shuyang Gu、Dong Chen、Jianmin Bao、Fang Wen、Bo Zhang、Dongdong Chen、Lu Yuan 和 Baining Guo。《用于文本到图像合成的向量量化扩散模型》。载于 IEEE/CVF 计算机视觉与模式识别会议论文集，第

10696 10706, 2022 3, 12, 13 – 页。[13] Xiangteng He 和 Yuxin Peng。《细粒度视觉文本表征学习》。IEEE视频技术电路与系统汇刊，30(2):520–531，2019年。5 [14] Jack Hessel、Ari Holtzman、Maxwell Forbes、Ronan Le Bras 和 Yejin Choi。《CLIPScore：用于图像描述的无参考评估指标》。载于 2021 年自然语言处理实证方法会议论文集，第 7514–7528 页。计算语言学协会，2021 年。6[15] Martin Heusel、Hubert Ramsauer、Thomas Unterthiner、Bern- hard Nessler 和 Sepp Hochreiter。《采用双时间尺度更新规则训练的生成对抗网络收敛于局部纳什均衡》。神经信息处理系统进展，30，2017。6[16] Tobias Hinz、Stefan Heinrich 和 Stefan Wermter。《生成式文本到图像合成的语义对象准确度》。IEEE模式分析与机器智能汇刊，2020 年。6[17] Jonathan Ho、Ajay Jain 和 Pieter Abbeel。《去噪扩散概率模型》。神经信息处理系统进展，33:6840–6851，2020 年。1、3[18] Jonathan Ho、Chitwan Saharia、William Chan、David J Fleet、Mohammad Norouzi 和 Tim Salimans。《用于高保真图像生成的级联扩散模型》。机器学习研究杂志，23:47–1，2022 年。1[19] Jonathan Ho 和 Tim Salimans。《无分类器扩散引导》。arXiv 预印本 arXiv:2207.12598，2022 年。5、11 [20] Susung Hong、Gyuseong Lee、Wooseok Jang 和 Seungry- ong Kim。《使用自注意力引导提升扩散模型的样本质量》，2022 年。5、12[21] Jin-Hwa Kim、Yunji Kim、Jiyoung Lee、Kang Min Yoo 和 Sang-Woo Lee。《互信息散度：多模态生成模型的统一度量指标》。arXiv 预印本 arXiv:2205.13445，2022 年。6[22] Doyup Lee、Chiheon Kim、Saehoon Kim、Minsu Cho 和 Wook-Shin Han。《Draft-and-revise：利用上下文 RQ 变换器实现有效的图像生成》。arXiv 预印本 arXiv:2206.04452，2022 年。1、3、6、8

[23] José Lezama、Huiwen Chang、Lu Jiang 和 Irfan Essa。《利用 token-critic 改进掩码图像生成》。载于欧洲计算机视觉会议，第 70–86 页。Springer，2022 年。2、3、6[24] Tsung-Yi Lin、Michael Maire、Serge Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Dollár 和 C Lawrence Zitnick。《Microsoft COCO：上下文中的常见物体》。载于欧洲计算机视觉会议，第 740–755 页。Springer，2014年。5[25] Alex Nichol、Prafulla Dhariwal、Aditya Ramesh、Pranav Shyam、Pamela Mishkin、Bob McGrew、Ilya Sutskever 和 Mark Chen。《GLIDE：利用文本引导扩散模型实现照片级真实感图像生成与编辑》。arXiv 预印本 arXiv:2112.10741，2021 年。1、6、11[26] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。《通过自然语言监督学习可迁移视觉模型》。载于国际机器学习会议，第 8748–8763 页。PMLR，2021 年。5[27] Alec Radford、Karthik Narasimhan、Tim Salimans 和 Ilya Sutskever。《通过生成式预训练改进语言理解》。2018。2[28] Aditya Ramesh、Prafulla Dhariwal、Alex Nichol、Casey Chu 和 Mark Chen。《利用 CLIP 潜变量进行分层文本条件图像生成》。arXiv 预印本 arXiv:2204.06125，2022 年。1[29] Aditya Ramesh、Mikhail Pavlov、Gabriel Goh、Scott Gray、Chelsea Voss、Alec Radford、Mark Chen 和 Ilya Sutskever。《零样本文本生成图像》。载于国际机器学习会议，第 8821–8831 页。PMLR，2021 年。1、2[30] Ali Razavi、Aaron Van den Oord 和 Oriol Vinyals。《利用 VQ-VAE-2 生成多样化的高保真图像》。神经信息处理系统进展，32，2019 年。2[31] Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Björn Ommer。《利用潜空间扩散模型进行高分辨率图像合成》。载于 IEEE/CVF 计算机视觉与模式识别会议论文集，第 10684–10695 页，2022 年。1、3[32] Chitwan Saharia、William Chan、Saurabh Saxena、Lala Li、Jay Whang、Emily Denton、Seyed Kamyar Seyed Ghasemipour、Burcu Karagol Ayan、S Sara Mahdavi、Rapha Gontijo Lopes 等。《具有深度语言理解能力的照片级真实感文本到图像扩散模型》。arXiv 预印本 arXiv:2205 11487，2022 年 1、3、11。。[33] Axel Sauer、Katja Schwarz 和 Andreas Geiger。《StyleGAN-XL：将 StyleGAN 扩展至大型多样化数据集》。载于 ACM SIGGRAPH 2022 会议论文集，第 1–10 页，2022 年。1[34] Ravi Shekhar、Sandro Pezzelle、Yauhen Klimovich、Aur-élie Herbelot、Moin Nabi、Enver Sangineto 和 Raffaella Bernardi。《Foil it! 找出图像与语言图注之间的一处不匹配》。arXiv 预印本 arXiv:1705.01359，2017。6[35] Yang Song、Jascha Sohl-Dickstein、Diederik P Kingma、Ab-hishek Kumar、Stefano Ermon 和 Ben Poole。《通过随机微分方程进行基于分数的生成建模》。arXiv 预印本 arXiv:2011.13456，2020 年。3

[36] Zhicong Tang、Shuyang Gu、Jianmin Bao、Chen Dong 和 Fang Wen。《改进的向量量化扩散模型》。arXiv 预印本 arXiv:2205.16007，2022年。1、2、3、5、6、11、16 [37] Ming Tao、Hao Tang、Songsong Wu、Nicu Sebe、Xiao-Yuan Jing、Fei Wu 和 Bingkun Bao。《DF-GAN：用于文本到图像合成的深度融合生成对抗网络》。arXiv 预印本 arXiv:2008.05865，2020。2、6[38] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。《注意力机制即一切》。神经信息处理系统进展，30，2017。2[39] Tao Xu、Pengchuan Zhang、Qiuyuan Huang、Han Zhang、Zhe Gan、Xiaolei Huang 和 Xiaodong He。《AttnGAN：利用注意力生成对抗网络进行细粒度文本到图像生成》。载于 IEEE 计算机视觉与模式识别会议论文集，第1316–1324页，2018。2、6[40] Jiahui Yu、Xin Li、Jing Yu Koh、Han Zhang、Ruoming Pang、James Qin、Alexander Ku、Yuanzhong Xu、Jason Baldridge 和 Yonghui Wu。《利用改进的 vqgan 进行向量量化图像建模》。arXiv 预印本 arXiv:2110.04627，2021。1、6[41] Jiahui Yu、Yuanzhong Xu、Jing Yu Koh、Thang Luong、Gunjan Baid、Zirui Wang、Vijay Vasudevan、Alexander Ku、Yinfei Yang、Burcu Karagol Ayan 等。《扩展自回归模型以实现内容丰富的文本到图像生成》。arXiv 预印本 arXiv:2206.10789，2022。1、2[42] Han Zhang、Tao Xu、Hongsheng Li、Shaoting Zhang、Xiaogang Wang、Xiaolei Huang 和 Dimitris N Metaxas。《StackGAN：利用堆叠生成对抗网络进行文本到照片级真实感图像合成》。载于 IEEE 国际计算机视觉会议论文集，第5907–5915页，2017。2 [43] Han Zhang、Tao Xu、Hongsheng Li、Shaoting Zhang、Xiaogang Wang、Xiaolei Huang 和 Dimitris N Metaxas。《StackGAN++：利用堆叠生成对抗网络实现逼真图像合成》。IEEE模式分析与机器智能汇刊，41(8):1947–1962，2018。2 [44] Yufan Zhou、Chunyuan Li、Changyou Chen、Jianfeng Gao 和 Jinhui Xu。《Lafite2：少样本文本到图像生成》。arXiv 预印本 arXiv:2210.14124，2022。2[45] Yufan Zhou、Ruiyi Zhang、Changyou Chen、Chunyuan Li、Chris Tensmeyer、Tong Yu、Jiuxiang Gu、Jinhui Xu 和 Tong Sun。《迈向文本到图像生成的无语言训练》。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第17907–17917页，2022。2、6 – 。[46] Minfeng Zhu、Pingbo Pan、Wei Chen 和 Yi Yang。《DM-GAN：用于文本到图像合成的动态记忆生成对抗网络》。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第5802–5810页，2019年。2

A. 实现细节 A.1. 引导采样训练



图9： **使用未采用无分类器引导训练的 TCTS 生成的样本（FID-30K：27.13，MID-L：12.34）**。

无分类器引导 [19] 是影响文生图扩散模型图像质量的关键因素 [25, 32], , 而且也已成功应用于基于令牌的模型 [5, 36]。由于无分类器引导用于推理阶段，因此训练过程中也必须使用它。然而，在训练期间确定引导尺度是一个难题。我们发现，引导尺度过高或过低都会损害 TCTS 的训练过程。TCTS 的训练流程从对真实图像中的随机词元进行掩码开始。随后，固定的生成器重建图像，而 TCTS 则经过训练以找出原先被掩码的位置。然而，由于较高的引导尺度会增强重建能力，TCTS 反而难以找到被掩码的位置，并倾向于输出平滑分布。另一方面，重建性能较差的较低引导尺度能为模型提供多样且容易的样本，从而使学习过程更快、更稳定。然而，由于实际推理时使用较高的引导尺度，模型的性能会非常低。样本见图9。因此，我们在训练流程中随机采样引导尺度，以此对任务难度进行正则化。这种引导尺度控制方法能够稳定训练、提升性能，并支持在推理时采用各种引导尺度设置。

A.2. 频率自适应采样

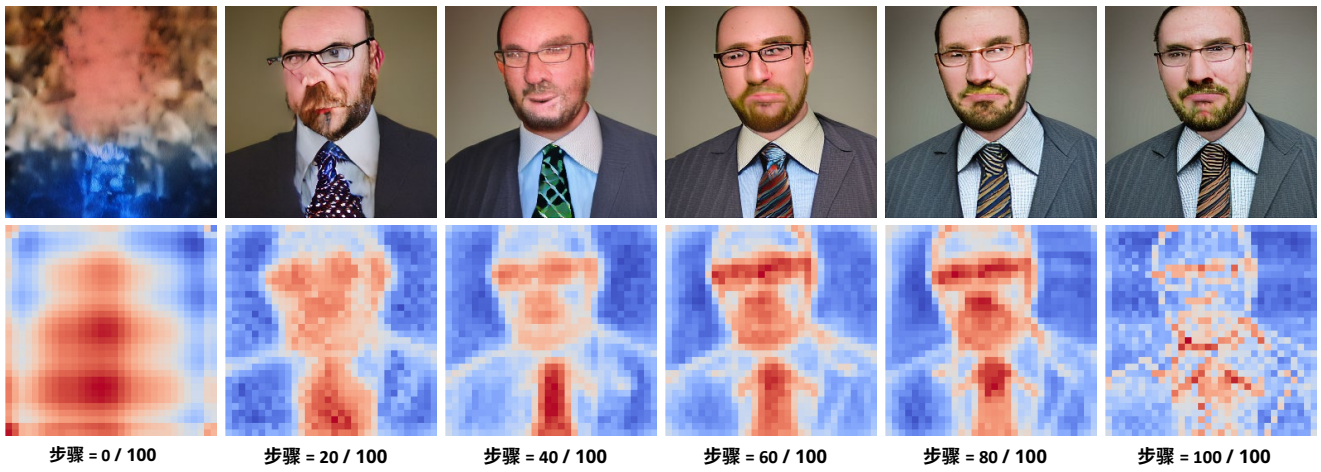


图10：**自注意力图的可视化**。**上方**：每个步骤的重建图像。**下方**：每个步骤的自注意力图可视化。

高度详细信息出现的区域 图 10在高频范围内具有较大的值。观察发现，随着生成过程的推进，所有区域的数值都逐渐降低，尤其是在简化区域。变化最先出现在优先进行简化的区域。在我们的 FAS 方法中，我们将该区域



图 11：**自注意力图的可视化，带阈值。上方**：合成图像。**下方**：自注意力掩码的可视化。($\phi = 0.5$)

按阈值进行划分，其可视化结果如 图 11所示。这类似于 [20]中所示的对象掩码或频率掩码。详细算法见 算法2。

算法 2频率自适应采样

输入： w**：持久权重，** map_{sa}**：自注意力图，** ϕ**：自注意力阈值，** G_θ**：生成器，** D_γ**：TCTS 模型，** c**：文本条件嵌入 1：** $\hat{x}_0, map_{sa} = G_\theta(x_t, c)$ **2：** $map_{tc} = D_\gamma(\hat{x}_0, c)$ **←TCTS 概率图 3：** $A_t = \{i | x_t^i =$**[掩码]**$\}$**4：** $L_t = \{i | map_{sa}^i < \phi\}$ **←低频位置 5：** $a = 1 + (w - 1) \times (n(A^C) \div N)$ **6：** **对于** $i \in A_t^C \cap L_t$ **do 7：** $map_{tc}^i = map_{tc}^i \times w$**8：结束循环 9：对于** $i \in A_t$ **do 10：** $map_{tc}^i = map_{tc}^i \times a$**11：结束循环 12：返回：** map_{tc}

此外，与 [20], 原始 Transformer 在 softmax 计算前对值应用 Softmax 函数不同，这里的自注意力图使用 Sigmoid 函数。

$$map_{sa}^{(h)} = sigmoid(Q_t^{(h)}(K_t^{(h)})^T/\sqrt{d}) \tag{2}$$

$$map_{sa} = GAP(map_{sa}^{(h)}) \tag{3}$$

这是因为 VQ 扩散模型 [12] 使用了 1024 的较大嵌入维度，导致 Softmax 函数的值之和会因位置不同而大幅下降。如果只将持久权重乘以低频位置，而不执行上文第 5 行和第 10 行所示的过程，那么对应低频位置的词元在采样过程中更有可能在概率上保留下来，从而可能无意中阻碍对象的生成。因此，在高频位置乘以权重“a”，有助于维持低频词元与高频词元的比例，同时产生将持久权重设为 1 的效果。

A.3. 超参数设置

在我们的实验中，超参数设置为：自注意力阈值（ϕ）设为 0.45，持久权重设为 15。对于可学习的 TCTS 模型，我们采用了 VQ 扩散 [12], 中的相同架构；但在使用 COCO 数据集时，将层数从 19 减少到 16；在使用 CUB 数据集时，将隐藏嵌入维度从 1024 减少到 512。

B. 附加样本

B.1. 过度简化样例



图12：**采用可撤销调度和较长推理步数的过度简化样例。顶部**：较长扩散步数（100步）下的图像生成过程。**底部**：较长扩散步数下的无掩码对象编辑，将“田野里一只伸着舌头的**狗** ”改为“田野里一只伸着舌头的**熊**”

B.2. 无掩码对象编辑



图13：**采用和未采用交叉注意力图加权采样的无掩码对象编辑样本**。从左侧图像开始，展示掩码比例为30%时每编辑20步后的结果图像。**顶部**：未采用加权采样的失败案例；**底部**：采用加权采样的结果。

在无掩码对象编辑中，以较低的掩码比例将大型物体进行转换颇具挑战，例如将甜甜圈变成西兰花。这是因为不同物体的词元分布截然不同；即使部分区域被掩码，原始物体的周围词元仍会影响生成结果。此外，对整幅图像重新采样可能导致背景等非必要部分发生显著变化。为了解决这一问题，我们提出利用交叉注意力图，提高对目标物体周围区域进行采样的权重，从而尽量减少对背景的不必要重新采样，使物体编辑更加容易。

B.3. 图像细化



图14：**增加细化步骤后的随机图像样本**。**上**：使用均匀采样 进行16步采样得到的原始图像，**中**：添加随机噪声后的精炼图像，**下**：使用TCTS得到的精炼图像。



图15：**通过掩码得分最低的词元进行细化得到的随机图像样本**。**上**：使用均匀采样 进行16步采样得到的原始图像，**下**：通过掩码TCTS中得分最低的词元得到的精炼图像。如前所述，由于该图像是使用相同的均匀采样方法重新生成的，因此难以确认图像质量是否有明显改善。不过，FID和MID方面的性能有所提升。

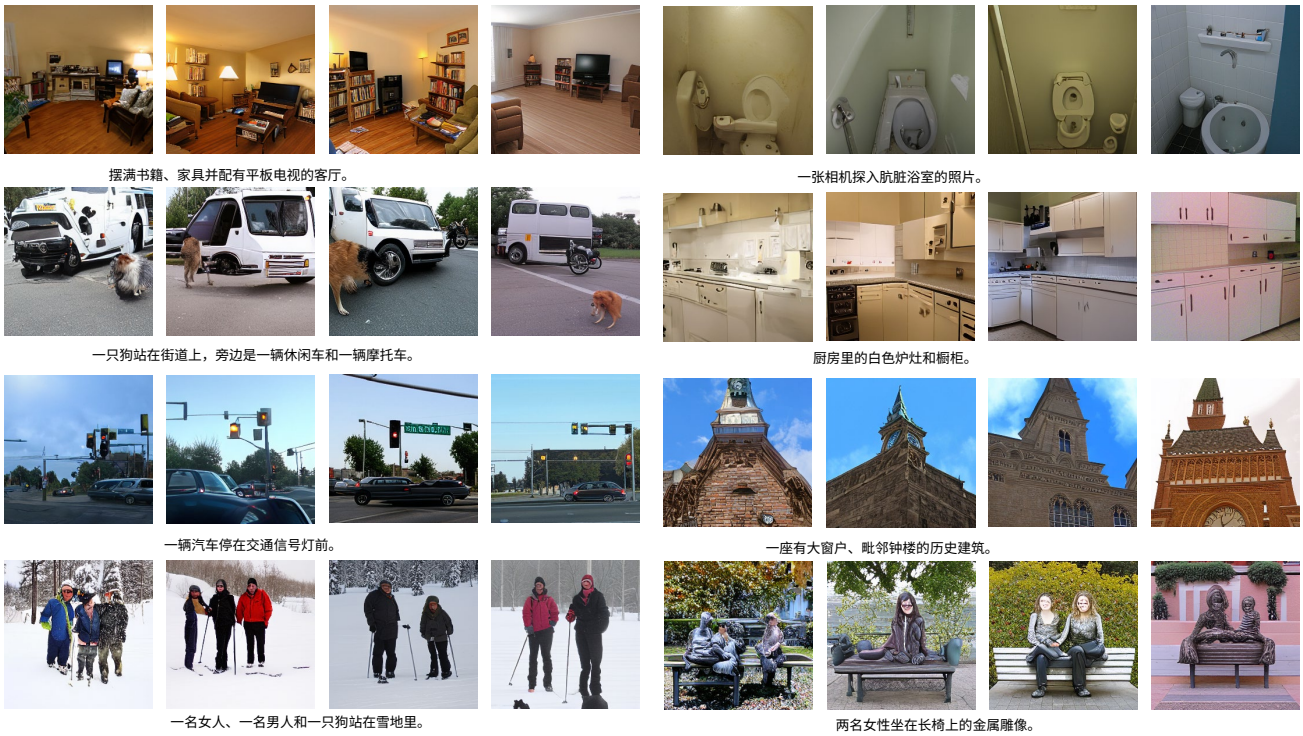


图16：**使用 TCTS 生成的样本。每段文本分别生成四张图像，扩散步数为 8、16、25 和 50 步。**



图17：**使用 TCTS 和 FAS 经 16 步生成的样本。**

C. 结果的进一步分析

C.1. 随时间变化的性能图

图 18 表明，在与文本相关的指标 MID 和 CLIP 分数方面，TCTS 优于其他基线方法。RR、TCTS 和 FAS 在 SOA 上的表现优于 Uniform，这为我们的分析提供了证据：可撤销方法提供了更多恢复机会，从而能够重新生成缺失对象。此外，该图还展示了 FAS 的影响：它显著提升了 TCTS 的 FID，同时保持了图像与文本之间的对齐。

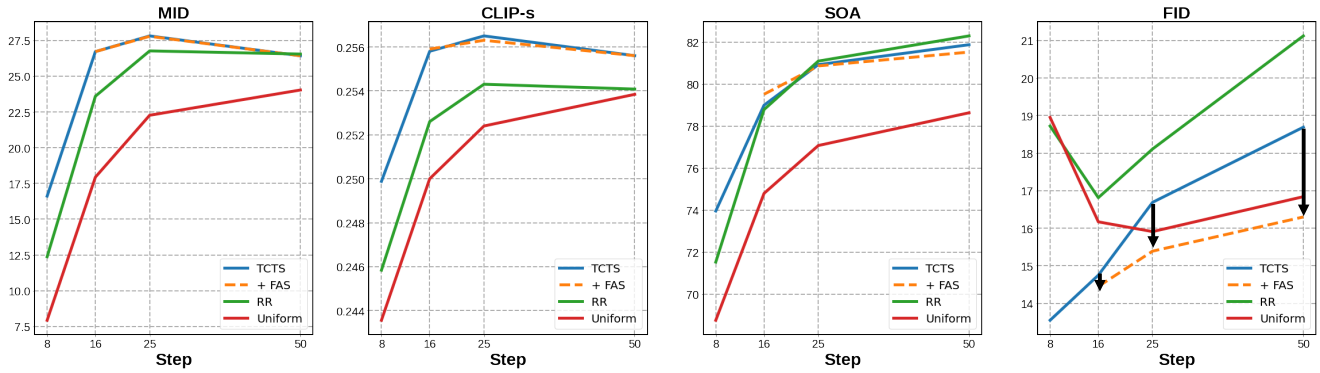


图18：**不同扩散步数下各方法的性能比较**。在我们的实验中，我们将无分类器引导固定为 5。使用 FAS 方法时，可以在保持文本对齐的同时降低 FID 分数。

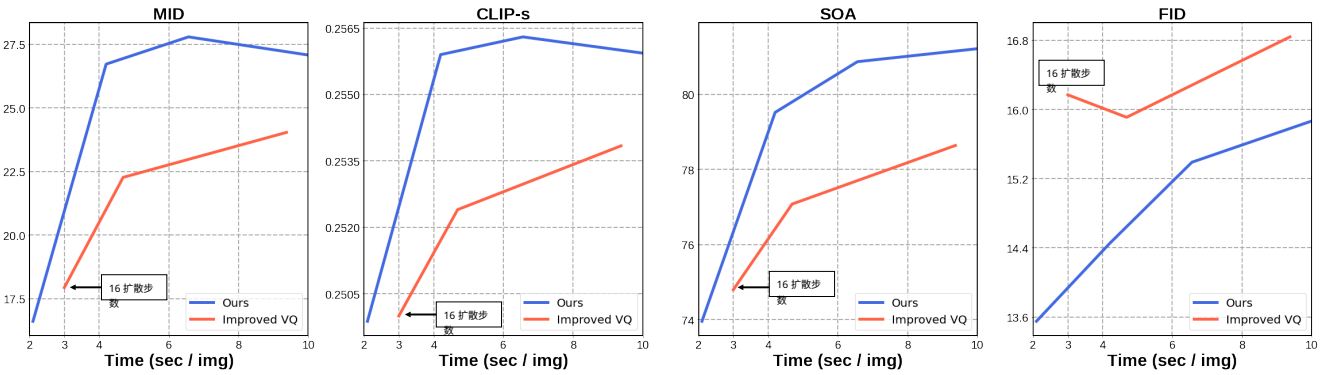


图19：**我们的模型与基线在生成时间上的性能比较**。在我们的实验中，我们将无分类器引导固定为 5。

我们将最终模型与基线改进型 VQ-Diffusion 进行了速度和质量评估 [36]。基线方法在 16 个扩散步数下生成一张图像需要 3 秒。正如 图 19所示，我们的模型在保持相同生成时间的同时，在所有指标上均优于基线。