TTS-VAR：一种用于视觉自回归生成的测试时缩放框架

Zhekai Chen[1] **Ruihang Chu**[2*] **Yukang Chen**[3] **Shiwei Zhang**[2] **Yujie Wei**[2]
Yingya Zhang[2] **Xihui Liu**[1*]

[1] HKU MMLab [2] Tongyi Lab, Alibaba Group [3] CUHK
zkchen66@outlook.com

Abstract

将视觉生成模型扩展到实际应用内容创作中是至关重要的，但这需要大量的训练和计算成本。作为替代方案，由于资源效率高且表现出色，测试时刻扩展受到了越来越多的关注。在这项工作中，我们提出了 TTS-VAR ，这是第一个视觉自回归（VAR）模型通用测试时刻扩展框架，将生成过程建模为路径搜索问题。为了动态平衡计算效率与探索能力，我们首先在因果生成过程的整个过程中引入了自适应降序批量大小调度。此外，受 VAR 的层次化由粗到细多尺度生成的启发，我们的框架整合了两个关键组件：(i) 在粗尺度中，我们观察到生成的标记很难进行评估，可能导致错误接受较劣样本或拒绝较优样本。注意到粗尺度包含足够的结构信息，我们提出了基于聚类的多样性搜索。它通过语义特征聚类保留结构多样性，从而在具有更高潜力的样本上进行后续选择。(ii) 在细尺度中，基于重采样的潜力选择通过潜力评分来优先考虑有前途的候选者，潜力评分被定义为结合多尺度生成历史的奖励函数。在强大的 VAR 模型 Infinity 上的实验显示了显著的 8.7 % GenEval 评分提高（0.690.75）。关键见解表明早期阶段的结构特征有效地影响了最终质量，以及重采样效果在不同生成尺度上存在差异。代码可在 https://github.com/ali-vilab/TTS-VAR 获得。

1 引言

近年来，图像生成模型取得了显著进展 [1–4] 。以前的文本到图像生成模型主要依赖于扩散模型 [5, 6] ，通过迭代去噪潜在变量从随机噪声中生成高质量图像。然而，大型语言模型（LLMs） [7–9] 的进步激发了对图像生成中自回归（AR）架构的兴趣 [10–12] ，利用顺序建模来捕捉视觉模式。在这些方案中，视觉自回归建模（VAR） [13, 14] 已经成为一个突破性的范式。它将图像编码为多尺度从粗到细的表示，并通过层次聚合逐步预测“下一个尺度”来合成图像。由于其优越的效率以及与 LLMs 统一集成的潜力，VAR 正迅速成为一个关键的研究前沿。

与此同时，随着 LLMs [15–19] 中测试时缩放的成功，研究人员已经开始探索将这种方法应用于图像生成模型以获得更好的结果。在自回归模型中，先前的工作通常将图像生成设计为一个图像级 [20] 或令牌级 [21] 的多阶段过程，将阶段序列视为思维链（CoT）。然而，这些方法需要额外的训练以实现有效的缩放。或者，基于扩散的方法 [22–26] 将缩放视为路径搜索问题，对不同的中间状态评分并选择最有前途的噪声进行去噪以获得更高质量的图像。实现显著改进的主要有两种策略。其一 [22, 23] 是引入额外的去噪步骤以获得用于图像解码的干净潜在变量，并根据最终解码结果选择中间潜在状态。另一种 [24] 则是直接从中间噪声潜在变量中对解码后的图像进行评分，通过奖励函数 [27] 来指导选择。

* 通讯作者

Preprint. Under review.

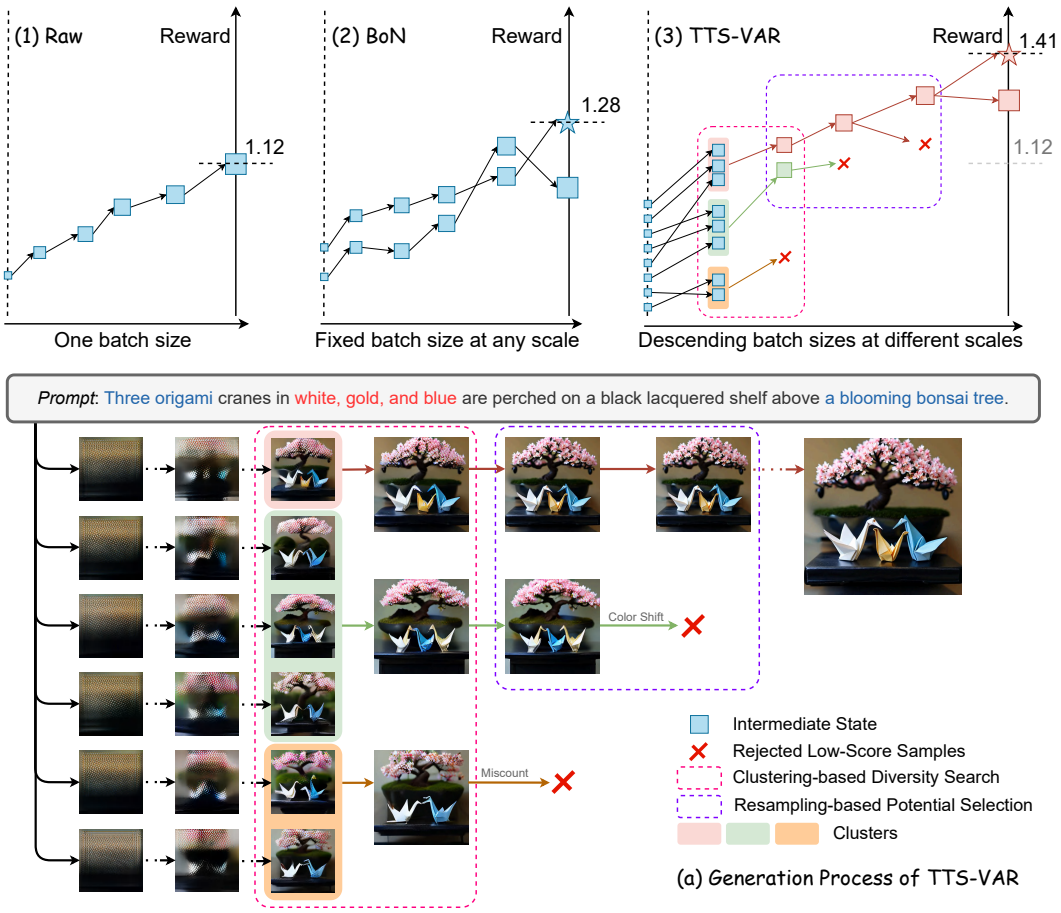


Figure 1: TTS-VAR 像 Best-of-N (BoN) 一样同时生成多个样本。在 TTS-VAR 中，我们采用自适应递减批量大小计划以最大限度地利用 AR 效率，早期阶段的特征聚类确保多样性，后期阶段根据潜力重新采样以获得更有价值的样本。(1-3) 是显示原始推理、BoN 和 TTS-VAR 之间差异的概述。(a) 是我们方法生成过程的详细示例。

受到这种观点的启发，我们探讨 VAR 模型是否也能从路径搜索中获益。然而，直接采用扩散模型中的两种策略并不简单。对于前者来说，额外的推理步骤会导致 VAR 模型的计算复杂度呈指数增长，从而带来极高的计算成本。这些额外步骤还会破坏 KV 缓存机制 [28, 29]，该机制对于在 AR 推理中保持效率至关重要。后者也未能达到预期。我们观察到，早期尺度的图像奖励难以准确代表最终图像的质量，导致错误地排除了一些可能在后期尺度生成中有潜力的早期尺度标记，如在 Sec. 5.3 中所示。我们将其归因于 VAR 和扩散模型之间的差异。在 VAR 中，不像扩散过程可以通过迭代去噪来优化生成的噪声，一旦生成，所有标记都保持不变。每个标记不仅对最终图像的解码有贡献，还直接影响所有后续标记的生成，导致对早期阶段的劣质标记的容忍度要低得多。

在本文中，我们介绍了用于 VAR 的第一个测试时刻缩放框架，简称 TTS-VAR 。与简单根据奖励函数进行选择不同，我们设计了与 VAR 的粗到细生成过程相匹配的缩放策略。首先，注意到 VAR 中 FLOPs 和 RAM 的消耗逐渐增加，我们在一个自适应的下降批量大小计划下实现了我们的框架，从粗略尺度中的较大批量大小缩小到细尺度中的较小批量大小。这在几乎没有额外消耗的情况下促进了更多可能性的表达。其次，尽管早期尺度难以通过奖励函数进行评估，但我们观察到结构信息对图像内容有很大影响，可以从早期尺度中捕捉到，如 Fig. 1 （b）和 **??** 所示。这促使我们将生成过程拆分为两个关键组成部分：用于早期尺度的基于聚类的多样性搜索和用于后期尺度的基于再采样的潜力选择。在早期尺度中，尽管难以估计中间状态的结果，我们的目标是在批量大小减小时保持多样性，从而在稍后选择潜力更高的样本。我们对由预训练提取器（如 DINOv2 [30]）提取的语义特征进行聚类，并从每个类别中挑选不同的样本以确保采样的多样性。在后期尺度中，由于中间图像的评分与最终图像的评分具有高度一致性，我们计算潜力评分以直接重新采样偏好样本。潜力评分是基于所有尺度的生成历史而不是仅当前尺度特定定义的奖励函数。

总的来说，我们提出了 TTS-VAR ，这是一种用于 VAR 模型的首个通用测试时缩放框架。通过整合基于聚类的多样性搜索和基于重采样的潜力选择，专门适应 VAR 的因果生成过程，TTS-VAR 始终提供稳定的性能提升。我们在经过扩展的文本到图像 VAR 模型 Infinity [14] 上进行了全面的实验和分析，揭示了为什么重采样方法表现出与规模相关的限制，并展示了结构特征聚类的优势。值得注意的是，TTS-VAR 显著提高了 GenEval 得分，从 0.69 提升至 0.75，同时在其他指标上也有稳定的提升。

2 相关工作

2.1 扩散模型中的测试时刻缩放

扩散模型 [6, 31–35] 通过将高斯分布去噪为图像分布来创建高分辨率图像。最初的研究工作 [36, 5, 37] 主要集中在增加去噪步骤的数量以提高图像质量。然而，研究表明，随着推理步骤数量的增加，性能达到了平台期，并且额外的采样步骤无效。因此，早期研究 [38–40] 主要旨在减少推理步骤的同时保持图像质量。

Ma 等人 [22] 将扩散模型中的缩放问题视为路径搜索问题，通过在潜在空间中应用若干搜索策略，并使用奖励函数作为验证器，取得了显著的改进。在此问题定义的基础上，后续研究 [23, 24] 探讨了各种搜索策略的有效性以及精确验证中间状态以供选择的方法。例如，大岛等人 [23] 使用了少步采样代替一步采样，以便获得更清晰且更适合验证的去噪图像。

2.2 自回归模型中的测试时缩放

在自回归大型语言模型 [9, 7, 8] 中，测试时缩放是一种广泛使用的技术，用于提高性能。自从 Wei 等人 [15] 提出连锁思维并使大型语言模型能够从结构化的思维过程中受益以来，各种研究 [16–19, 41] 探索了树搜索、图搜索和其他方法，以进一步改善结果。所有这些策略都利用了模型的推理能力，并利用了自然语言中固有的性质进行缩放。

然而，在自回归图像生成模型中，由于其特征是具有稳定令牌长度的确定性过程，直接将图像令牌序列作为“思考”增加是不自然的。相反，Guo 等人 [20] 将生成 CoT 概念化为一个图像级问题。通过采用一个统一的生成和理解模型 [42] ，该模型首先生成然后评估，自我纠正结果以符合预期。然而，这种方法仅依赖于评估结果，忽视了生成过程本身。Jiang 等人 [21] 提出将任务分为语义层次和令牌层次阶段，从而使多阶段生成成为思考过程。然而，这种方法需要额外的强化学习来进行微调。

3 预备知识：视觉自回归建模

与传统的下一个标记预测自回归模型不同，如 LLama-Gen [10] ，视觉自回归建模（VAR）[13] 将输入图像 I 标记化为特征图 $F \in \mathbb{R}^{h\times w\times d}$ 。通过量化器 Q ，它将特征图 F 量化为多尺度离散残差特征图序列 [43] $\{r_i\}_{i=1}^{K}$ ，其中 K 表示在不同分辨率下的残差特征的数量。对于每个残差特征图 r_k ，分辨率为 $h_k \times w_k$ ，它逐步从 $k=1$ 增加到 $k=K$ 。具体而言，当 $k=1$ 时，$h_k = w_k = 1$ ，当 $k=K$ 时，$h_k = h, w_k = w$ 。从残差特征的序列中，在每个尺度 k ，可以计算出一个逐渐精炼的特征图 f_k 为：

$$f_k = \sum_{i=1}^{k} \text{up}(r_i, (h, w)), \tag{1}$$

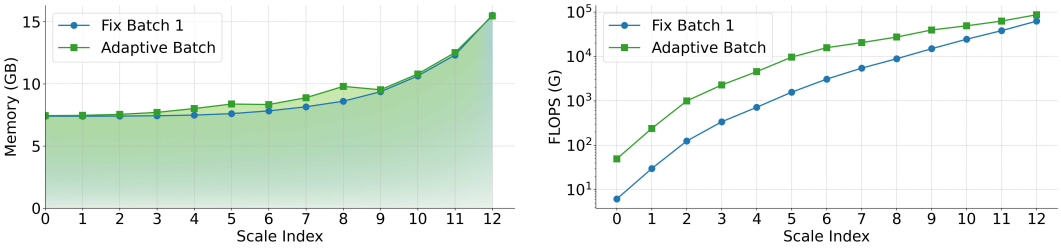
其中 up$(\cdot,\cdot)$ 表示将单尺度特征图上采样到目标分辨率，并且 f_k 是特征 $\{r_i\}_{i=1}^{k}$ 的聚合和。在推理过程中，下采样的累计特征图 $\tilde{f}_k = \text{down}(f_k, (h_{k+1}, w_{k+1}))$ 被附加为预测下一个尺度的初始标记，并使用尺度化的因果遮罩以便于局部双向信息建模。该变压器被训练用于预测下一尺度的残差特征图。在 Infinity [14] 中，一个基于 VAR 的模型被扩展用于文本到图像生成，量化器 Q 从 VQ [44] 升级为 BSQ [45] 。另外，使用 Flan-T5 [46] 文本编码器 Ψ 进行提示嵌入。以文本提示 c 为条件，总体似然为：

$$p(r_1, r_2, \ldots, r_K) = \prod_{k=1}^{K} (r_k | r_1, r_2, \ldots, r_{k-1}; \Psi(c)). \tag{2}$$

4 方法

在 TTS-VAR 中，我们将生成高质量且人类偏好的图像概念化为路径搜索问题，并确定两个主要子问题：(i) 如何搜索更多的可能性，以及 (ii) 如何选择中间状态以获得更优的最终结果。除了应用如 Sec. 4.1 中所示的自适应批量大小计划来扩大搜索范围外，我们在 Sec. 4.2 中引入基于聚类的多样性搜索，并在 Sec. 4.3 中采用基于重采样的潜在选择来解决这些问题。完整的方法如 Fig. 1 (c) 所示。

4.1 自适应批采样



(a) Memory Usage (b) Computational Complexity

Figure 2: 不同的批大小安排。我们在 (a) 中可视化了内存使用情况，在 (b) 中可视化了生成 Infinity 的 13 个比例尺的计算复杂度，分别是固定批大小为 1 和自适应批大小。具体来说，这里的自适应批大小是 [8,8,6,6,6,4,2,2,2,1,1,1,1]。这种批大小安排使得在很少额外消耗的情况下能够实现更多的可能性。

受因果注意力机制的影响，在 VAR 模型中，RAM 内存消耗和计算费用都会随着标记序列长度的增加而增加。如 蓝色 中的 Fig. 2 (a) 和 (b) 所示，在早期尺度的推理过程中，内存需求和计算成本是最小的。然而，在后期尺度中，当已有的序列显著延长且当前预测尺度包含大量标记时，资源消耗变得相当大。

因此，我们在推理过程中实施自适应批量大小，利用早期尺度效率较低的特点。这个递减的自适应批量大小计划为 $\{b_0, b_1, \ldots, b_K\}$，其中 K 表示尺度的数量，在早期阶段生成更多样本，在后期阶段生成较少样本。对于包含 13 个尺度的典型 VAR 模型，批量大小计划为 $\{8N, 8N, 6N, 6N, 6N, 4N, 2N, 2N, 2N, 1N, 1N, 1N, 1N\}$，除非另有说明。在早期尺度，Sec. 4.2 中的聚类会过滤多个类别。在后期尺度，Sec. 4.3 中的重采样选择出更优的状态。如 Fig. 2 中所示，尽管增大的批次数在早期尺度增加了内存和计算成本，但这些增加在总体开销中是相对较小的。

4.2 基于聚类的多样性搜索

随着序列长度的增加，保持大批量就变得成本过高，因此需要一种筛选方法来选取所需样本。一种简单的方法是计算中间结果的奖励函数，并选择得分较高的那些。然而，我们在 Sec. 5.3 中的研究发现，评分模型难以对早期中间图像进行一致于最终图像的奖励评分评估，这一点也被 Guo 等人 [20] 观察到。为了避免错误地淘汰可能在后期生成中具有潜力的某些初始样本，我们探索保留样本多样性的方法。

我们观察到，在生成过程中，结构信息与后期尺度中出现的细节不同，它显著影响最终图像的质量，并从早期阶段开始传递。我们在 **??** 中分析了这一现象，发现诸如 DINOv2 [30] 之类的提取器可以有效捕捉与结构紧密相关的特征。基于此，我们根据语义特征创建集群，以过滤每个类别的样本，并确保结构的多样性，从而最大限度地提高获得有价值结果的可能性。

具体来说，在当前批大小为 b_i 的情况下，我们需要选择 b_{i+1} 个样本作为下一个状态。首先，对于 b_i 个中间图像 $\{I_j\}_{j=1}^{b_i}$，每个图像通过特征提取器 F 嵌入到高维语义嵌入空间中，生成一组嵌入 $S = \bigcup_{b_i} F(I_j)$。随后，我们应用 K-Means++ [47] 算法将这些嵌入聚类为 b_{i+1} 个聚类中心，并选择与聚类中心 L2 距离最短的样本作为新批次。

为了提取关于多样性的结构信息，我们主要采用自监督 DINOv2 [30] 作为提取器，生成特征图 $s \in \mathbb{R}^{(h' \times w') \times d}$。为了获得用于聚类的一维特征，我们对特征块 $s' \in \mathbb{R}^{(h' \times w')}$ 应用了 PCA 降维。我们还考虑对 $s' \in \mathbb{R}^d$ 的第二维进行池化，以及使用 InceptionV3 [48] 上的监督特征，但不包括最终的全连接层。我们在 **??** 中讨论了这些选择。

4.3 基于重采样的潜在选择

与通过聚类进行早期阶段的多样性保持相反，当中间图像与最终结果显示出高度一致性时，奖励函数可以在后期直接引导生成向更高质量和与人类偏好一致的方向发展。通常，一个奖励函数 $r_\phi(x)$ 是从奖励模型 ϕ 派生出来的，其中包括专门训练的评分模型和视觉-语言模型。对于基于文本提示 c 的得分，奖励函数可以表示为 $r_\phi(x, c)$ 。在基于生成分布 $p_\theta(x)$ 的生成模型背景下，我们旨在引导分布与奖励偏好对齐 [49, 24] ，如下所示：

$$p_{\theta'}(x) = \frac{1}{Z} p_\theta(x) \exp(\lambda \cdot r_\phi(x, c)) \tag{3}$$

，其中 $p_{\theta'}(x)$ 是目标分布，Z 是归一化常数，λ 是用于控制选择中的温度的超参数。

为了获得高质量的样本，我们评估当前尺度 k 下每个中间状态的奖励分数，并用基于潜在分数 P_k 的多项分布中采样的状态替换它们。考虑到 VAR 的生成是一个带有历史状态的路径，仅评估由图像解码器 $\mathcal{D}$ 从累积特征 f_k 解码的图像 $x_k = \mathcal{D}(f_k)$ 可能无法充分反映最终结果的潜力，因此我们思考几种潜在分数。

潜在分数 P_k 。我们将 $P_k(x_0, x_1, \ldots, x_k)$ 表示为 k 尺度下样本的潜在分数，其中 $x_0, x_1, \ldots, x_k$ 表示该样本的生成历史。

- $P_k(x_0, x_1, \ldots, x_k) = \exp(\lambda \cdot r_\phi(x_k, c))$ ： 这种方法直接利用奖励分数作为潜在分数，称为价值。这也被称为重要性采样（IS） [50, 51] 。
- $P_k(x_0, x_1, \ldots, x_k) = \exp(\lambda \cdot (r_\phi(x_k, c) - r_\phi(x_{k-1}, c)))$ ： 这计算两个连续尺度之间的差异作为潜在分数，称为 DIFF。
- $P_k(x_0, x_1, \ldots, x_k) = \exp(\lambda \cdot \max_{i=0}^{k}\{r_\phi(x_i, c)\})$ ： 这选择生成路径中的最高分作为当前潜在分数，称为 MAX。
- $P_k(x_0, x_1, \ldots, x_k) = \exp(\lambda \cdot \sum_{i=0}^{k} r_\phi(x_i, c))$ ： 这将累积来自历史的所有分数以确定当前的潜在分数，标记为 SUM。

不同的潜在评分有利于生成历史的不同属性。例如，DIFF 偏好具有较高增长率的样本，而 MAX 偏好具有较高天花板的样本。在我们的设置中，VALUE 表现良好。我们将在 Sec. 5.3 中探讨这些选择。

5 实验

在本节中，我们展示了 TTS-VAR 在强大的 VAR 模型 Infinity-2B [14] 上具有重采样温度 $\lambda = 10$ 的有效性。我们在 Sec. 5.1 和 Sec. 5.2 中展示了比较，并在 Sec. 5.3 和 **??** 中精确分析设计细节。按照之前的工作 [24, 22, 52] ，我们使用 ImageReward [53] 作为指导的奖励函数。我们使用主要指标 GenEval [54] 和 T2I-CompBench [55] 进行结果评估，相关指标有 ImageReward [53] 、HPSv2.1 [56, 57] 、Aesthetic V2.5 [58] 和 CLIP-Score [59, 60] ，这些指标基于 GenEval 提供的提示。

5.1 整体性能

如表 1 所示，我们提出的方法在现有最先进的模型和传统的测试期间扩展策略（重要性采样和最佳 N 值）上表现出显著改进。模型规模为 2B 参数的情况下，TTS-VAR 在 $N = 8$ 处实现了总体 GenEval 分数 0.7530，超过了 Stable Diffusion 3（8B 参数）的 0.74 记录，同时使用了少 60 % 的参数。我们的框架在单个项目上也表现出显著提升，比如两个物体。值得注意的是，即便计算开销很小（$N = 2$ ），我们的方法仍然取得了 0.7403 的竞争性能，超越了使用 25 % 样本数量的最佳 N （$N = 8$ ）。

从不同 N 的角度来看，TTS-VAR 在 GenEval 和 ImageReward 指标中始终在不同样本量下保持一致的性能提升，如图 3 所示。虽然 Best-of-N 抽样从较大的 N 中获益更多，但它的性能明显不如我们的，即使在 $N = 2$ 和 $N = 8$ 时也未能达到我们的结果。在附录中提供了不同 N 值的详细评估，以供综合分析。

我们进一步评估了在 T2I-CompBench [55] 上的性能。如所展示的，TTS-VAR 在每个指标上相比于基础模型都有显著的提升。与在 GenEval 上的结果一致，我们的方法在 $N = 2$ 上取得了比 Best-of-N 在 $N = 8$ 上更优异的结果，并在大多数单项和整体平均分上获得了最高分。

Methods	# Params	GenEval ↑			
		Two Obj.	Counting	Color Attri.	Overall
Diffusion Models					
SDXL [2]	2.6B	0.74	0.39	0.23	0.55
+FK ($N = 8$) [24]	2.6B	-	-	-	0.65
PixArt-Alpha [1]	0.6B	0.50	0.44	0.07	0.48
DALL-E 3 [3]	-	0.87	0.47	0.45	0.67
FLUX [4]	12B	0.81	0.74	0.45	0.66
SD3 [61]	8B	0.94	0.72	0.60	0.74
AR Models					
LlamaGen [10]	0.8B	0.34	0.21	0.04	0.32
Chameleon [62]	7B	-	-	-	0.39
Show-o [42]	1.3B	0.52	0.49	0.28	0.53
+PARM ($N = 20$) [20]	1.3B	0.77	0.68	0.45	0.67
Emu3 [11]	8.5B	0.71	0.34	0.21	0.54
Infinity	2B	0.8351	0.5923	0.6150	0.6946
Infinity+IS ($N = 8$)	2B	0.8969	0.6220	0.6550	0.7181
Infinity+BoN ($N = 8$)	2B	0.9201	0.6756	0.6700	0.7364
Infinity+Ours ($N = 2$)	2B	0.9278	0.7113	0.6775	0.7403
Infinity+Ours ($N = 8$)	2B	0.9501	0.7411	0.6800	0.7530

Table 1: Quantitative evaluation on GenEval .

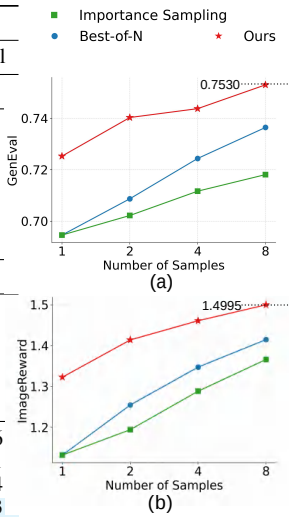


Figure 3: 得分曲线随样本数量 N 的变化

Model	Avg.	Color	Shape	Texture	2D Spatial	3D Spatial	Numeracy	Non-spatial	Complex
		B-VQA	B-VQA	B-VQA	UniDet	UniDet	UniDet	S-CoT	S-CoT
Stable v2 [31]	0.4839	0.5065	0.4221	0.4922	0.1342	0.3230	0.4582	0.7567	0.7783
Stable XL [2]	0.5255	0.5879	0.4687	0.5299	0.2133	0.3566	0.4988	0.7673	0.7817
Pixart- α -ft [1]	0.5583	0.6690	0.4927	0.6477	0.2064	0.3901	0.5032	0.7747	0.7823
DALLE · 3 [3]	0.6168	0.7785	0.6205	0.7036	0.2865	0.3744	0.5926	0.7853	0.7927
FLUX.1 [4]	0.6087	0.7407	0.5718	0.6922	0.2863	0.3866	0.6185	0.7809	0.7927
Infinity [14]	0.5688	0.7421	0.4557	0.6034	0.2279	0.4023	0.5479	0.7820	0.7890
Infinity+IS ($N = 8$)	0.5965	0.7746	0.5078	0.6501	0.2462	0.4194	0.6002	0.7803	0.7937
Infinity+BoN ($N = 8$)	0.6115	0.7950	0.5439	0.6886	0.2545	0.4205	0.6090	0.7870	0.7937
Infinity+Ours ($N = 2$)	0.6151	0.7887	0.5578	0.6858	0.2697	0.4286	0.6112	0.7853	0.7936
Infinity+Ours ($N = 8$)	0.6230	0.8073	0.5914	0.7121	0.2644	0.4302	0.6340	0.7880	0.7963

Table 2: Quantitative evaluation on T2I-CompBench .

5.2 定性比较

我们在这里展示了由不同方法生成的图像以供参考，以进一步说明图像质量和文本对齐的改进。如 Fig. 4 所示，我们的方法正确生成了所需数量的物体，例如，在第一个例子中是“皮划艇”，在第二个例子中是“鸟”。这是基础模型和其他缩放策略难以实现的。在包含多个数字对和颜色的复杂场景中，TTS-VAR 确保了在第四个例子中“金色苹果”之类的颜色属性，避免了属性遗忘的问题。

5.3 利用重采样获取更优样本

分析。在路径搜索中，使用具有高潜力的中间状态进行重采样是一种简单但有效的方法，可以以最小消耗获得优异的结果。然而，在 VAR 中，这可能并不有利，甚至在某些尺度上可能有害。在 Fig. 5 (a) 中，我们展示了仅使用 Best-of-N ($N = 2, 4$) 和在特定尺度上同时应用潜力（VALUE）重采样之间的分数差异。尽管 Best-of-N 选择确保了比较高的结果，但很明显，在早期尺度（例如，尺度 3）进行重采样会导致最终结果显著下降。相反，在后期尺度进行重采样会带来一定程度的改进。

我们从中间状态与最终图像之间一致性的角度分析这一现象，如 Fig. 5 (b) 所示。我们在每个尺度上计算潜在分数，并相应地选择潜力最高的一个。然后，我们评估所选的最佳中间状态是否与最佳最终图像（尺度 12 的最佳状态）一致，从而得到一个 0 到 1 之间的分数序列，称之为一致性。在这个演变曲线的早期尺度中，一致性较低表明这些分数很少能准确反映最终结果的质量。从某个比较晚的尺度开始，例如尺度 6，分数具有相对较高的一致性，因此变得有价值。这解释了为什么再采样的效果会有所不同，并且应选择性地应用于后期尺度上。

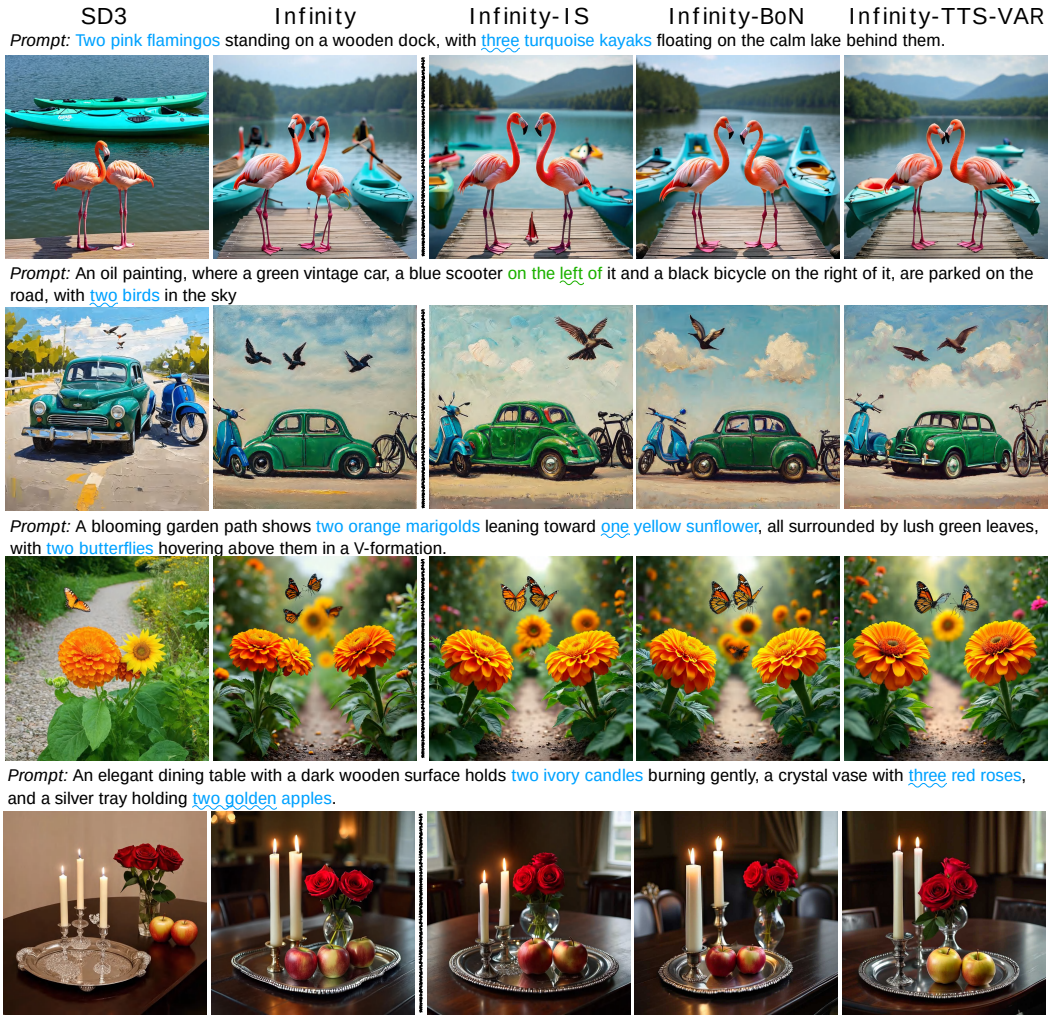


Figure 4: 定性比较。每一行显示了由稳定扩散 3 (SD3) [61] 、Infinity 和使用测试时间缩放策略的 Infinity 生成的结果，其中对象标记为蓝色，关系标记为绿色。

重采样尺度。基于前述观察，我们进一步研究在后期阶段提高重采样频率是否有利。如 Table 3 所示，与原始推断相比，重采样极大地增强了结果。然而，频率的增加几乎没有影响。例如，与尺度 [6,9] 相比，在尺度 [6,8,10] 上 ImageReward 和 HPS 略有提高，但 Geneval 有所损失。考虑到执行重采样会带来与图像解码和评分计算相关的额外计算开销，我们选择只在尺度 6 和 9 进行重采样。

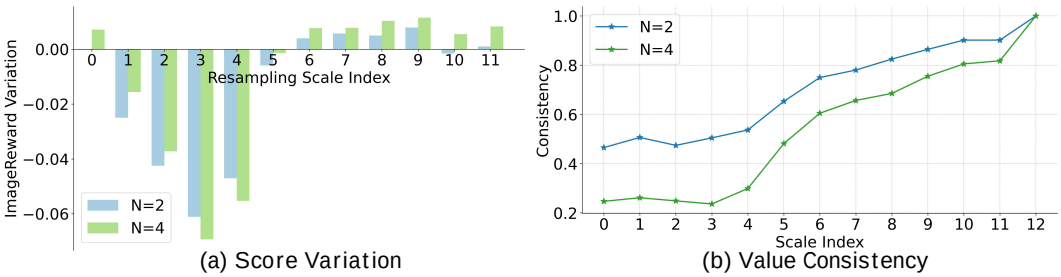


Figure 5: 重采样选择。左图显示了在不同尺度（0-11）下执行基于重采样的潜在选择时，ImageReward 分数的变化。右图显示了中间状态分数与每个尺度下最终结果分数的一致性。这表明在 VAR 中，并非所有尺度都适合选择，有些甚至可能导致退化。

N	Resampling Scale	GenEval	ImageReward	HPS	CLIP	Aesthetic
1	-	0.6946	1.132	0.3042	0.3366	0.5811
2	[6, 9]	0.7133	1.2572	0.3066	0.3379	0.5801
2	[6, 8, 10]	0.7130	1.2591	0.3066	0.3381	0.5809
2	[6, 7, 8, 9, 10, 11]	0.7114	1.2497	0.3067	0.3378	0.5810
4	[6, 9]	0.7276	1.3534	0.3082	0.3398	0.5817
4	[6, 8, 10]	0.7247	1.3592	0.3085	0.3397	0.5822
4	[6, 7, 8, 9, 10, 11]	0.7210	1.3558	0.3083	0.3398	0.5830

Table 3: 重采样尺度差异。此表显示了不同重采样尺度的结果，表明了重采样频率的影响。

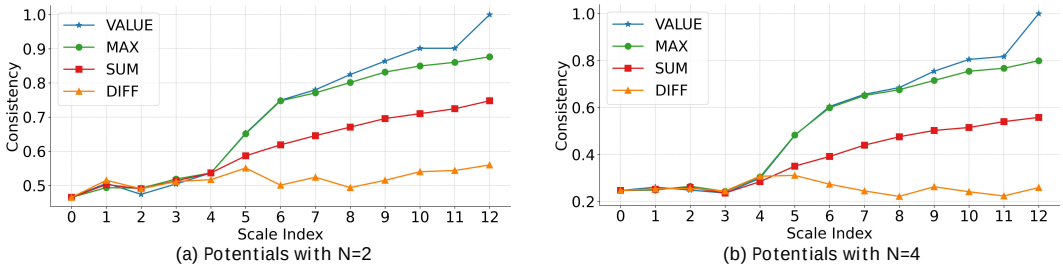


Figure 6: 不同潜力的一致性。我们可视化了不同潜力评分与最终结果之间的一致性。因此，VALUE 和 MAX 能够更好地指示潜力。

潜在分数。在 Sec. 4.3 中提到，我们已经开发了不同的计算模式来探索那些具有更高潜力的对象。最初，我们通过理论方法来寻找更好的选择，方法是可视化一致性。如所展示的，DIFF 连续呈现低一致性水平，未能预测出预期的结果。SUM 表现出稳定的增幅，但数值相对较低。相反，VALUE 和 MAX 展示了相似的特征，自 6 级起保持相对较高的得分，并显示出稳定的增加。

在 Table 4 中的实验与 $N = 2, 4$ 呈现出一致的结果。在这些潜在选项中，DIFF 在所有指标上表现滞后。尽管 SUM 取得了一些可接受的结果，但总体分数仍然较低。如预期的一致性所预测，VALUE 和 MAX 在与文本相关的指标中，如 GenEval、ImageReward 和 HPS，取得了最高分，表明选择出优异最终结果的可能性更大。考虑到 MAX 需要在每个尺度上进行评分计算，并导致额外的计算成本，我们采用 VALUE 作为潜在评分。

分析。关于 Sec. 5.3 ，再采样并不是对所有尺度普遍适用。然而，在 VAR 中，与早期尺度相关的效率和低成本提供了一个无价且不可错过的机会，以便寻找更多样本，从而释放出最终结果的更大潜力。我们注意到，对于相同的提示，图像的结构显著影响评分。此外，与后期出现的细节不同，结构信息可以从早期尺度中捕获。Fig. 7 的右侧表明生成过程遵循从结构到细节的进程，从尺度 2 开始可以识别出粗略的轮廓。当使用 DINOv2 提取中间图像并通过 PCA 进行可视化时，如底线所示，这些特征表现出类似于原始图像的特性。因此，我们利用结构信息并进行基于聚类的多样性搜索，以采样不同结构，从而扩大更多可能性，特别是在再采样可能不够时。

聚类尺度。据 Fig. 7 ，尺度 2 的特征显示出粗略的结构，而尺度 5 的特征揭示了类似于最终结果的精细结构。因此，我们特意在这些尺度上应用了聚类。带有和不带有聚类的 $N = 2, 4$ 的结果展示在 Table 5 中。每个模块的第一行表示没有应用聚类（Best-of-N）的结果，后续行展示了在不同聚类尺度下的结果。显然，每一次聚类都增加了产生更好结果的可能性，并且在同时使用这两个尺度时有明显的增长。

特征提取器。我们测试了 Table 6 中概述的用于聚类特征的各种提取器。PCA 和 Pool 都是由 DINOv2 [30] 提取的特征的变换，如 Sec. 4.2 中详细描述。虽然监督学习的 InceptionV3 [48]

N	Potential	GenEval	ImageReward	HPS	CLIP	Aesthetic
2	VALUE	0.7133	1.2572	0.3066	0.3379	0.5801
2	MAX	0.7150	1.2510	0.3065	0.3379	0.5803
2	SUM	0.7130	1.2364	0.3064	0.3379	0.5801
2	DIFF	0.7006	1.1725	0.3042	0.3365	0.5798
4	VALUE	0.7276	1.3534	0.3082	0.3398	0.5817
4	MAX	0.7285	1.3495	0.3082	0.3398	0.5815
4	SUM	0.7244	1.3326	0.3082	0.3394	0.5830
4	DIFF	0.7030	1.2412	0.3051	0.3378	0.5808

Table 4: 潜在得分差异。此表显示使用不同潜在计算方式作为重采样得分的结果。

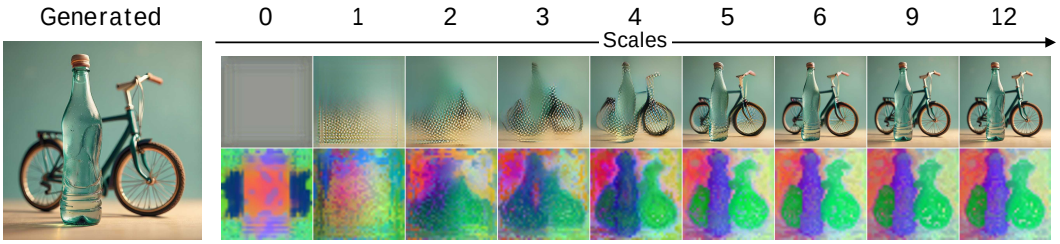


Figure 7: 生成过程的可视化。这里的文本提示是“一张瓶子和自行车的照片”。左边是一张最终生成的图像。右边是相应的生成过程和从不同尺度提取的可视化 DINOv2 特征。这表明，在早期尺度捕获的特征可以指示结构信息。

N	Clustering Scale	GenEval	ImageReward
2	-	0.7087	1.2545
2	[2]	0.7089	1.2513
2	[5]	0.7099	1.2558
2	[2, 5]	0.7184	1.2682
4	-	0.7244	1.3471
4	[2]	0.7300	1.3502
4	[5]	0.7293	1.3558
4	[2, 5]	0.7337	1.3610

Table 5: 聚类尺度差异。此表展示在特定尺度上进行和不进行聚类的结果。

N	Extractor	GenEval	ImageReward
2	PCA	0.7184	1.2682
2	Pool	0.7127	1.2720
2	Inception	0.7073	1.2727
4	PCA	0.7337	1.3610
4	Pool	0.7296	1.3629
4	Inception	0.7207	1.3664

Table 6: 特征提取差异。此表展示了采用不同特征提取方法时的结果。这里的 PCA 和 Pool 都是由 DINOv2 提取的二维特征转换而来的。

特征在 ImageReward 中表现最优，但在 GenEval 中表现明显不佳。PCA 平均提供了更优的结果，且被采用。我们将此归因于 PCA 的 patch-level 特征更紧密地与观察到的结构特征相一致。

6 结论

在这项工作中，我们介绍了第一个用于 VAR 模型的一般测试时间缩放框架。通过对不同尺度的分析，我们证明了 TTS-VAR ，其中结合了自适应批量采样、基于聚类的多样性搜索以及基于重采样的潜在选择，与 VAR 生成过程的不同阶段相一致。这种双策略方法在保持算法效率的同时，以极少的额外计算成本提高了最终结果质量。我们注意到了隐私和版权方面的限制和潜在的社会影响，并在附录中进行了讨论。

7

附录

A TTS-VAR 的算法

我们在 Alg. 1 中描述了 TTS-VAR 的算法。遵循 VAR 生成过程 [13]（Infinity [14]），TTS-VAR 首先预测当前尺度下的残差标记并将其添加到累积特征图中。在需要聚类的尺度下，TTS-VAR 使用提取器从特征图解码出的 b_i 中间图像中收集特征。然后，它根据这些特征对样本进行聚类，并选择 b_{i+1} 个样本作为下一批。在需要重新采样的尺度下，TTS-VAR 使用潜在函数计算每个图像的评分，并从多项分布中抽样 b_{i+1} 个索引以获得更优的中间状态。

Algorithm 1 TTS-VAR

Require: Scales $\mathcal{S} = \{s_1, s_2, ..., s_K\}$, Descending batch sizes $\mathcal{B} = \{b_1, b_2, ..., b_K\}$, Clustering scales set S_c , Resampling scales set S_r , Generative model θ , Reward model r_ϕ Extractor F , Potential Score function P , Text prompt c .

Initialize accumulated feature map f_0 with zeros.
for $i \in \{1, 2, ..., K\}$ **do** ▷ Iterate through scales
  $r_i \leftarrow \text{Generate}(\theta, b_i, s_i, f_{i-1}, c)$
  $f_i \leftarrow f_{i-1} + r_i$
  if $s_i \in S_c$ **then** ▷ Clustering phase
    $x \leftarrow \text{Decode}(f_i)$
    $feat \leftarrow F(x)$
    $index \leftarrow \text{KMeans++}(feat, b_{i+1})$
    $f_i \leftarrow f_i[index]$
  else if $scale \in S_r$ **then** ▷ Resampling phase
    $x \leftarrow \text{Decode}(f_i)$
    $rw \leftarrow r_\phi(x)$
    $p \leftarrow P(rw)$
    $index \leftarrow \text{Multinomial}(p, b_{i+1})$
    $f_i \leftarrow f_i[index]$
  end if
end for
return Final generated images $\text{Decode}(f_K)$

B 详细的主要结果

N	Strategy	GenEval	ImageReward	HPS	CLIP	Aesthetic
1	Raw Inference	0.6946	1.1320	0.3042	0.3366	0.5811
1	Ours	0.7253	1.3226	0.3084	0.3395	0.5822
2	Importance Sampling	0.7022	1.1941	0.3051	0.3374	0.5807
2	Best-of-N	0.7087	1.2545	0.3069	0.3384	0.5813
2	Ours	0.7403	1.4136	0.3106	0.3411	0.5821
4	Importance Sampling	0.7116	1.2883	0.3067	0.3387	0.5815
4	Best-of-N	0.7244	1.3471	0.3083	0.3397	0.5820
4	Ours	0.7437	1.4605	0.3112	0.3414	0.5821
8	Importance Sampling	0.7181	1.3657	0.3085	0.3395	0.5810
8	Best-of-N	0.7364	1.4144	0.3103	0.3406	0.5820
8	Ours	0.7530	1.4995	0.3122	0.3420	0.5810

Table 6: 不同策略的得分。本表展示了不同缩放策略的结果。

我们在 Table 6 中展示了变体曲线的详细结果。显而易见，TTS-VAR 在所有指标 [54, 53, 57, 58, 60] 上相较于基线显示出明显的优势。在 Table 7 中，我们列出了 GenEval [54] 指标的每一项。总体而言，我们的方法显著提高了处理两个物体和计数任务的性能。我们将此归因于多角色场景中结构准确性的重要性，尤其是在涉及两个物体和多个相同物体（计数）时。举例来说，当提供三个物体的提示时，模型可能错误地生成一个包含四个物体的布局。一旦出

现此错误，按照 VAR 中的从结构到细节的生成过程，后续尺度很难纠正。然而，TTS-VAR 有助于结构多样性，从而能够选择正确配置的布局，避免对劣质样本产生不可逆转的错误生成过程。

N	Strategy	Overall	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.
1	Raw Inference	0.6946	0.9938	0.8351	0.5923	0.9293	0.2020	0.6150
1	Ours	0.7253	0.9938	0.9072	0.6518	0.9192	0.2096	0.6700
2	Importance Sampling	0.7022	0.9969	0.8497	0.6071	0.9268	0.1869	0.6475
2	Best-of-N	0.7087	0.9906	0.8789	0.6339	0.9242	0.1944	0.6300
2	Ours	0.7403	0.9936	0.9278	0.7113	0.9318	0.1995	0.6775
4	Importance Sampling	0.7116	0.9906	0.8840	0.6339	0.9318	0.1970	0.6325
4	Best-of-N	0.7244	1.0000	0.8969	0.6756	0.9242	0.1944	0.6550
4	Ours	0.7437	0.9906	0.9510	0.6994	0.9293	0.2045	0.6875
8	Importance Sampling	0.7181	0.9906	0.8969	0.6220	0.9318	0.2121	0.6550
8	Best-of-N	0.7364	0.9938	0.9201	0.6756	0.9444	0.2146	0.6700
8	Ours	0.7530	0.9969	0.9501	0.7411	0.9318	0.2172	0.6800

Table 7: GenEval 详情。此表显示了 GenEval 基准测试的每个项目，“Object”是“Obj.”的缩写，“Attribute”是“Attri.”的缩写。

C 性能与计算消耗的关系

我们在此展示了随着计算量 Fig. 8 增加，GenEval、ImageReward 和 HPSv2 的变化曲线，以及样本数量 N 的增长。如图所示，我们的方法 TTS-VAR 具有更高的计算效率，并以不到一半的 TFLOPs 超过了重要性采样和最佳 N 选择法。

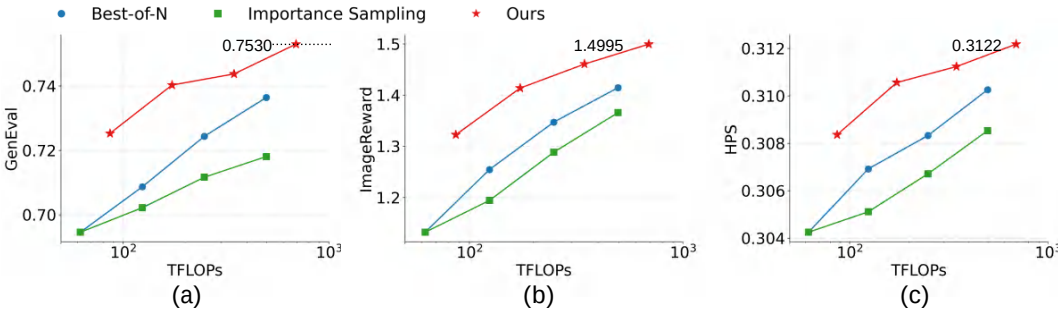


Figure 8: 性能相对于 Flops。这张图展示了不同方法的变异曲线，以计算消耗为横轴，展示了我们方法的效率。

D 消融研究

D.1 流水线消融

我们在 Table 8 中展示了总体流程中不同设计组件的消融研究。由于单独使用自适应批量采样（没有集成样本选择机制）无法直接提高生成性能，这些情况用“- ”表示。在排除这些基线情况后，基于聚类的多样性搜索和基于重采样的潜在选择都显示出性能的提升，在奖励和相关评估指标上观察到具有统计显著性的增长。

值得注意的是，聚类方法带来了相对适度的改进，这可以归因于其主要功能是维持结构多样性，而不是主动识别用于后续生成的优质样本。通过聚类维持多样性和通过重采样进行基于质量的选择的结合协同增强了整个流程的有效性。这个双机制框架最终在基准系统上实现了显著的性能提升，其中重采样组件在选择用于迭代优化的高质量候选者方面起到了关键作用。

D.2 奖励模型

我们对使用不同的奖励模型来评价中间图像并计算潜在分数（VALUE）进行了比较，包括 Aesthetic [58] 、ImageReward [53] 、HPSv2 [57] 和 HPS+ImageReward。由于 HPS 和

N	Method	GenEval	ImageReward	HPS	CLIP	Aesthetic
2	Infinity	0.6946	1.1320	0.3042	0.3366	0.5811
	+BoN	0.7087	1.2545	0.3069	0.3384	0.5813
	+Adaptive Batch Sampling	-	-	-	-	-
	+Clustering-Based Diversity Search	0.7220	1.2591	0.3072	0.3385	0.5816
	+Resampling-Based Potential Selection	0.7403	1.4136	0.3106	0.3411	0.5821
4	Infinity	0.6946	1.1320	0.3042	0.3366	0.5811
	+BoN	0.7244	1.3471	0.3083	0.3397	0.5820
	+Adaptive Batch Sampling	-	-	-	-	-
	+Clustering-Based Diversity Search	0.7294	1.3608	0.3095	0.3403	0.5824
	+Resampling-Based Potential Selection	0.7437	1.4605	0.3112	0.3414	0.5821

Table 8: 流程切除。本表显示了每个设计的收益。

ImageReward 的数值范围不同，对于 HPS+ImageReward，我们首先分别使用这两个模型计算分数，然后将每个模型的数值 softmax 到范围 $[0, 1]$ ，最后取平均值作为潜在分数。

如 Table 9 所示，通常，每个奖励模型都会促进相应指标的提高。例如，在 $N = 4$ 中，美学模型、ImageReward 模型和 HPS 模型分别在相关指标中获得了最高分。在不同的模型中，ImageReward 促进的改进更显著。尤其是在 $N = 2$ 中，ImageReward 在 GenEval 中表现出明显的领先地位，甚至在 HPS 分数中击败了 HPS。我们将这归因于其能够清晰地区分优劣样本，并且评分更符合人类偏好的能力。

N	Reward Model	GenEval	ImageReward	HPS	CLIP	Aesthetic
2	-	0.7087	1.2545	0.3069	0.3384	0.5813
2	Aesthetic	0.6966	1.123	0.3054	0.3366	0.6004
2	ImageReward	0.7403	1.4136	0.3106	0.3411	0.5821
2	HPS	0.7135	1.2246	0.3102	0.3391	0.583
2	HPS+ImageReward	0.7238	1.3522	0.3088	0.3402	0.5824
4	-	0.7244	1.3471	0.3083	0.3397	0.5820
4	Aesthetic	0.6842	1.1172	0.3056	0.3363	0.6114
4	ImageReward	0.7437	1.4605	0.3112	0.3414	0.5821
4	HPS	0.7255	1.2812	0.3154	0.3402	0.5843
4	HPS+ImageReward	0.7413	1.4128	0.3101	0.3406	0.5818

Table 9: 奖励模型消融。这张表格显示了使用不同模型对潜力的结果。

D.3 λ 设置

在 Table 10 中，我们展示了在固定聚类操作的情况下，使用不同温度 λ 进行重采样过程的结果。直观来说，更高的温度促进了具有更高潜在分数的中间状态的表达，并防止了较优样本的产生。然而，过高的温度也可能扩大具有最高分数的中间状态与那些分数仅略低的状态之间的差距。这可以直接抑制这些略微落后的中间状态的生成，而这些状态可能最终成为最佳结果。如图所示，尽管 ImageReward 稳步上升，但在 GenEval 中 $\lambda = 10.0$ 落后于 $\lambda = 5.0$ 与 $N = 4$ 。

N	Lambda	GenEval	ImageReward	HPS	CLIP	Aesthetic
2	-	0.7087	1.2545	0.3069	0.3384	0.5813
2	0.1	0.7065	1.2458	0.3065	0.3387	0.5811
2	0.5	0.7210	1.3167	0.3080	0.3400	0.5819
2	1.0	0.7222	1.3459	0.3087	0.3403	0.5821
2	5.0	0.7361	1.4010	0.3101	0.3410	0.5821
2	10.0	0.7403	1.4136	0.3106	0.3411	0.5821
4	-	0.7244	1.3471	0.3083	0.3397	0.5820
4	0.1	0.7308	1.3576	0.3089	0.3400	0.5816
4	0.5	0.7347	1.3918	0.3094	0.3406	0.5819
4	1.0	0.7418	1.4097	0.3099	0.3407	0.5820
4	5.0	0.7465	1.4500	0.3108	0.3412	0.5820
4	10.0	0.7437	1.4605	0.3112	0.3414	0.5821

Table 10: λ 消融。此表格显示了不同 lambda 值的结果。

E 更多可视化结果

我们在 Fig. 9 上展示了 GenEval 的结果对，以比较 Infinity、Infinity-IS、Infinity-BoN 和 Infinity-TTS-VAR 的质量。这些案例是从 GenEval 的文本提示中抽样的，包括单一对象、两个对象、

计数、颜色、位置和颜色属性。如图所示，我们的方法为单一对象生成了更高质量的样本，例如左侧第二行的飞机，有效地避免了生成伪影。在两个对象的设置中，TTS-VAR 成功区分了对不同对象的引用，并根据提示生成准确的输出。例如，在右侧第二行，它消除了概念混合和对象消失。正如在右侧第 3 到第 10 行所示，TTS-VAR 在确定计数数字、位置关系和颜色属性方面也表现出色。值得注意的是，在最后一行，我们的方法生成了一个违反直觉的“绿色胡萝卜”，展示了其将对象与其自然属性分离的能力。

F 社会影响

当应用于 VAR 模型时，TTS-VAR 增强了生成图像与文本描述的一致性，使生成过程更加可控，更适合满足创意和生产需求。然而，我们也认识到这种方法可能被滥用，从而导致隐私和版权问题。然而，我们相信，我们对 VAR 生成过程的深入研究将帮助研究人员获得更清晰的理解，推动对生成过程中可控性和安全性的研究，最终确保图像生成模型成为安全且可管理的工具。

G 局限性和未来工作

虽然 TTS-VAR 相比于基线显示出了显著的改进并创下了新记录，但它仍有两个主要限制。首先，TTS-VAR 没有完全解决文本提示与生成图像之间的错位问题。正如 Table 7 中的得分所示，仍然存在一些失败案例，特别是在 Position 这一项。其次，尽管 TTS-VAR 基于一个通用的由粗到细的过程，其在其他粗到细模型上的潜在应用仍未被探索，比如使用一维标记器的自回归模型。未来，我们将更深入地研究生成过程，分析失败的原因，并设计解决方案以释放进一步的扩展潜力。此外，我们计划评估 TTS-VAR 与其他自回归粗到细模型的兼容性，包括那些使用一维标记器和结合扩散模型的混合架构的模型。这些努力旨在为文本到图像合成创建一个更具鲁棒性的扩展框架，同时增强粗到细范式的方法学转移性。

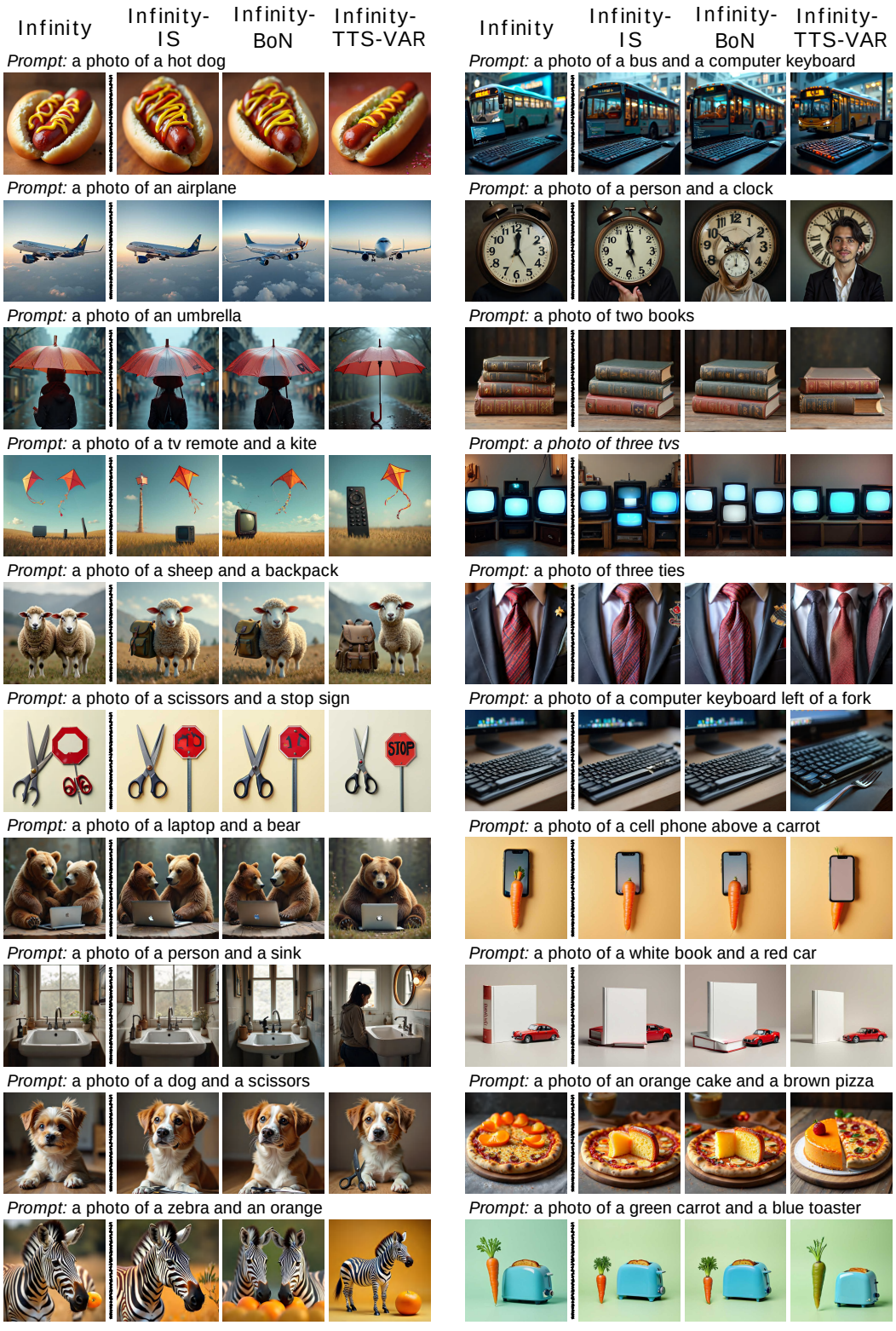


Figure 9: 更多的可视化结果。样本是从 GenEval 提示生成的，其中“IS”表示重要性采样，“BoN”表示最佳 N 选，而我们的方法为 TTS-VAR 。我们展示了各种情况，包括单个物体、两个物体、计数、颜色、位置和颜色属性。

References

[1] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv: 2310.00426*, 2023.

[2] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv: 2307.01952*, 2023.

[3] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

[4] Black Forest Labs. Flux. https://blackforestlabs.ai/announcing-black-forest-labs/, 2024.

[5] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv: 2010.02502*, 2020.

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NEURIPS*, 2020.

[7] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[8] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[9] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[10] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv: 2406.06525*, 2024.

[11] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv: 2409.18869*, 2024.

[12] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv: 2410.13848*, 2024.

[13] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Neural Information Processing Systems*, 2024. doi: 10.48550/arXiv.2404.02905.

[14] Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. *arXiv preprint arXiv: 2412.04431*, 2024.

[15] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, F. Xia, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Neural Information Processing Systems*, 2022.

[16] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[17] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[18] Maciej Besta, Nils Blach, Ale Kubíek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michal Podstawski, H. Niewiadomski, P. Nyczyk, and Torsten Hoefler. Graph of thoughts: Solving elaborate problems with large language models. *AAAI Conference on Artificial Intelligence*, 2023. doi: 10.1609/aaai.v38i16.29720.

[19] Ruomeng Ding, Chaoyun Zhang, Lu Wang, Yong Xu, Minghua Ma, Wei Zhang, Si Qin, Saravan Rajmohan, Qingwei Lin, and Dongmei Zhang. Everything of thoughts: Defying the law of penrose triangle for thought generation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 1638–1662. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.95.

[20] Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Peng Gao, Hongsheng Li, and Pheng-Ann Heng. Can we generate images with cot? let's verify and reinforce image generation step by step. *arXiv preprint arXiv: 2501.13926*, 2025.

[21] Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot. *arXiv preprint arXiv: 2505.00703*, 2025.

[22] Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, and Saining Xie. Inference-time scaling for diffusion models beyond scaling denoising steps. *arXiv preprint arXiv: 2501.09732*, 2025.

[23] Yuta Oshima, Masahiro Suzuki, Yutaka Matsuo, and Hiroki Furuta. Inference-time text-to-video alignment with diffusion latent beam search. *arXiv preprint arXiv: 2501.19252*, 2025.

[24] Raghav Singhal, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen McKeown, and Rajesh Ranganath. A general framework for inference-time scaling and steering of diffusion models. *arXiv preprint arXiv: 2501.06848*, 2025.

[25] Masatoshi Uehara, Yulai Zhao, Chenyu Wang, Xiner Li, Aviv Regev, Sergey Levine, and Tommaso Biancalani. Inference-time alignment in diffusion models with reward-guided generation: Tutorial and review. *arXiv preprint arXiv: 2501.09685*, 2025.

[26] Ye Tian, Ling Yang, Xinchen Zhang, Yunhai Tong, Mengdi Wang, and Bin Cui. Diffusion-sharpening: Fine-tuning diffusion models with denoising trajectory sharpening. *arXiv preprint arXiv: 2502.12146*, 2025.

[27] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *arXiv preprint arXiv: 2203.02155*, 2022.

[28] Yuhan Liu, Hanchen Li, Yihua Cheng, Siddhant Ray, Yuyang Huang, Qizheng Zhang, Kuntai Du, Jiayi Yao, Shan Lu, Ganesh Ananthanarayanan, Michael Maire, Henry Hoffmann, Ari Holtzman, and Junchen Jiang. Cachegen: Kv cache compression and streaming for fast large language model serving. *Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, 2023. doi: 10.1145/3651890.3672274.

[29] Luohe Shi, Hongyi Zhang, Yao Yao, Zuchao Li, and Hai Zhao. Keep the cost down: A review on methods to optimize llm' s kv-cache consumption. *arXiv preprint arXiv: 2407.18003*, 2024.

[30] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024, 2024.

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *arXiv preprint arXiv: 2112.10752*, 2021.

[32] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv: 2204.06125*, 2022.

[33] Enze Xie, Junsong Chen, Yuyang Zhao, Jincheng Yu, Ligeng Zhu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, Han Cai, Bingchen Liu, Daquan Zhou, and Song Han. Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. *arXiv preprint arXiv: 2501.18427*, 2025.

[34] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[35] Haonan Qiu, Shiwei Zhang, Yujie Wei, Ruihang Chu, Hangjie Yuan, Xiang Wang, Yingya Zhang, and Ziwei Liu. Freescale: Unleashing the resolution of diffusion models via tuning-free scale fusion. *arXiv preprint arXiv:2412.09626*, 2024.

[36] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *arXiv preprint arXiv: 2206.00364*, 2022.

[37] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv: 2011.13456*, 2020.

[38] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv: 2202.00512*, 2022.

[39] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv: 2303.01469*, 2023.

[40] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv: 2211.01095*, 2022.

[41] Jiankai Sun, Chuanyang Zheng, Enze Xie, Zhengying Liu, Ruihang Chu, Jianing Qiu, Jiaqi Xu, Mingyu Ding, Hongyang Li, Mengzhe Geng, et al. A survey of reasoning with foundation models: Concepts, methodologies, and outlook. *ACM Computing Surveys*, 2023.

[42] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv: 2408.12528*, 2024.

[43] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. *arXiv preprint arXiv: 2203.01941*, 2022.

[44] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6306–6315, 2017.

[45] Yue Zhao, Yuanjun Xiong, and Philipp Krähenbühl. Image and video tokenization with binary spherical quantization. *arXiv preprint arXiv: 2406.07548*, 2024.

[46] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020.

[47] Davin Choo, C. Grunau, Julian Portmann, and Václav Rozho. k-means++: few more steps yield constant approximation. *International Conference on Machine Learning*, 2020.

[48] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. *Computer Vision and Pattern Recognition*, 2015. doi: 10.1109/CVPR.2016.308.

[49] Tomasz Korbak, Ethan Perez, and C. Buckley. Rl with kl penalties is better viewed as bayesian inference. *Conference on Empirical Methods in Natural Language Processing*, 2022. doi: 10.48550/arXiv.2205.11275.

[50] V. Elvira and Luca Martino. Advances in importance sampling. *Wiley StatsRef: Statistics Reference Online*, 2021. doi: 10.1002/9781118445112.stat08284.

[51] Art Owen and Yi Zhou Associate and. Safe and effective importance sampling. *Journal of the American Statistical Association*, 95(449):135–143, 2000. doi: 10.1080/01621459.2000.10473909.

[52] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and S. Levine. Training diffusion models with reinforcement learning. *International Conference on Learning Representations*, 2023. doi: 10.48550/arXiv.2305.13301.

[53] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Neural Information Processing Systems*, 2023. doi: 10.48550/arXiv.2304.05977.

[54] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[55] Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47:3563–3579, 2025. doi: 10.1109/TPAMI.2025.3531907.

[56] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. *IEEE International Conference on Computer Vision*, 2023. doi: 10.1109/ICCV51070.2023.00200.

[57] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv: 2306.09341*, 2023.

[58] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv: 2210.08402*, 2022.

[59] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv: 2103.00020*, 2021.

[60] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *Emnlp*, 2021.

[61] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

[62] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv: 2405.09818*, 2024.