种子能保证可复现吗？GPU 上不能：文生图模型中的批次依赖型非确定性

Javier Conde、Xiuyuan Qi、Shanshan Liu 和 Pedro Reviriego

arXiv:2611.30545v1 [cs.AI] 23 Nov 2026

摘要——**在文生图 AI 模型中，人们普遍认为，固定提示词和随机种子就能固定生成的图像。我们证明，在 GPU 上并非如此。我们在三种 GPU 类型（RTX A6000、RTX 6000 Ada 和 H100 NVL）上，以四种数值精度和量化方式评估了六种文生图模型（扩散模型和自回归模型）；每种模型最多使用 100 条提示词。对于比较的每种配置，我们在提示词和随机种子固定的情况下，对比单独生成的图像与在批处理中生成的图像，也比较了不同 GPU 上的结果。本文报告的每项回放控制都达到位精确，因此这种变化并非随机噪声：执行设置，尤其是批次组成和 GPU，决定了输出。该效应显著，且因模型而异。对于一个采用 bf16 的自回归模型，单图输出与批处理输出之间的结构相似性降至 0.580，采用 int8 时则降至 0.551。对于第二个采用 bf16 的自回归模型，使用采样解码时该相似性降至 0.261，使用贪心解码时降至 0.229；对于一个 fp8 扩散模型，则降至 0.68。一个采用 bf16 的扩散模型生成的结果与批次1输出接近（0.96–0.98），但每张图像都发生了变化。在大多数情况下，批处理图像虽然不同，但同样有效；而在某些情况下，例如提示词包含数字和计数器时，图像的视觉效果明显更差。跨 GPU 的差异与批次效应处于同一量级：针对一个采用 bf16 的自回归模型，在两种 GPU 架构（RTX 6000 Ada 与 RTX A6000）之间进行比较时，没有任何图像逐位相同；跨 GPU 的学习感知图像块相似度（LPIPS）为 0.355 至 0.383，与同一 GPU 内的批次效应（0.347 至 0.381）相当。这种模式与批次依赖型内核执行相符：归约顺序会随张量形状变化，而 8 位格式会放大微小数值差异，使其变得明显。其影响是直接的：随机种子无法保证结果可复现；混合不同批次大小或硬件的基准测试和图像质量比较可能产生偏差；对生成图像进行安全审计和 A/B 测试时，需要报告服务配置。**

索引词—**文本到图像生成、可复现性、非确定性、批量推理、量化、GPU**。

I. 引言

TEXT-TO-IMAGE 模型以大规模方式提供服务。为了充分发挥 GPU 的效率，推理引擎会将请求分组成批处理；为了降低内存占用和成本，它们会以低精度运行权重和激活值，例如 bf16、fp8 或 int8。另一方面，用户和基准测试通常认为，只要固定提示词和随机种子，生成的图像就会固定。

这一假设在若干场景中都至关重要。图像质量基准测试通过模型的输出对其进行比较。审计和安全性评估根据生成图像得出结论。服务栈的 A/B 测试会比较变更前后的输出。所有这些比较都假设：在提示词和随机种子相同的情况下，输出中的任何差异都源于正在测试的变更，而非执行设置。图像编辑就是一个直接示例。编辑模型接到指令，要修改图像中的一个物体，同时保持其他所有内容不变[?], [?]。如果输出依赖于执行设置，那么当批次组成、精度或硬件发生变化时，同一项编辑请求可能会以不同方式保留未改动区域。因此，用户要求保留的区域可能会因服务配置而非指令本身而改变。由此，保留未编辑区域就成了依赖执行环境的生成在应用层面的影响，而不仅仅是数值细节。

对于大语言模型，这一假设已知并不成立。采用贪心解码并固定随机种子时，某个请求的输出可能会因批处理中其他请求的变化而改变，因为浮点加法不满足结合律，且归约顺序取决于批次形状 [1], [2], [3],[4]。目前已有多种解决方法，从批次不变内核到将请求的执行隔离或固定的调度方案 [3], [5], [6], [7], [8], [9], [10]。

对于文生图模型，人们对此知之甚少。Kim 等人发现，在 Stable Diffusion 中，相同的提示词和随机种子在不同 GPU 上会生成不同图像；这一研究是在生成内容验证和共识的背景下开展的 [11]。Fang 等人研究了扩散语言模型（一种文本模型系列）在受控因素下的非确定性 [12]。实践者报告称，生成图像会随批次大小而变化以及在广泛使用的工具中跨 GPU 的情况[1]。这些报告仅为轶闻，并未系统性地衡量其影响。

本文测量文生图中的批处理相关非确定性。我们固定提示词、随机种子和初始潜变量，并比较单独生成的图像与在批处理中生成的图像，涵盖不同模型、精度、批次大小和 GPU。据我们所知，这是首次针对扩散模型和自回归模型的文生图生成，对批次大小、量化以及硬件导致的非确定性进行系统性评估。

本文的贡献包括：

J. Conde 和 P. Reviriego 任职于西班牙马德里的马德里理工大学电信工程高等技术学院信息处理与电信中心（电子邮箱：javier.conde.diaz@upm.es；pedro.reviriego@upm.es）。X. Qi 和 S. Liu 任职于中国成都的电子科技大学（电子邮箱：xiuyuanqi@std.uestc.edu.cn；ssliu@uestc.edu.cn）。

[1]diffusers 问题 #2659 和 #6222，https://github.com/huggingface/diffusers/issues/2659，https://github.com/huggingface/diffusers/issues/6222；ComfyUI 问题 #788，https://github.com/comfyanonymous/ComfyUI/issues/ 788；AUTOMATIC1111 问题 #10847，https://github.com/AUTOMATIC1111/stable-diffusion-webui/issues/10847。

1) 一种运行协议，可将执行引起的变化与随机噪声区分开来；通过回放控制证明，相同运行的结果位精确匹配；并采用分阶段设计，设置包含 10、25 和 100 个提示词的嵌套提示集，因此，规模更大的运行是在先前结果的基础上扩展，而非取而代之。

2) 在三种 GPU 类型上，对来自文本到图像领域两个主要类别——扩散和自回归——的六种模型，在四种精度与量化设置下进行测量。

3) 使用四项指标（比特一致性、平均绝对误差、结构相似性指数 [13], 学习感知图像块相似度 [14]）量化这一影响，并纳入批处理图像明显更差、而不只是有所不同的最坏情况。4) 将这一影响归因于批处理、精度和 GPU。自回归模型所呈现的批处理效应大于我们测量的每一种扩散配置。我们讨论这对基准测试、审计和服务的影响。

本文其余部分安排如下。第二节回顾相关工作。第三节介绍模型、硬件、提示词、实验方案和指标。第四节展示结果。第五节讨论其影响和局限性，第六节作出总结。

II. 相关工作

A. 神经网络执行中的非确定性

浮点加法不满足结合律，因此并行归约的结果取决于各部分和的合并顺序。这一性质对高性能计算和深度学习工作负载可复现性的影响已有充分记录[15]。深度学习框架也引入了自身的非确定性来源，例如自动调优算法选择和原子操作，这些因素也列在框架文档中 [16], [17]。Gundersen 等人综述了机器学习中更广泛的不可复现性来源 [18], ，而 Summers 和 Dinneen 表明，在训练过程中，初始参数哪怕只有一位发生变化，也会导致最终模型截然不同 [19]。

针对大语言模型推理，Atil 等人测量了“确定性”设置在实践中无法保持确定性的情况 [1]，，Yuan 等人追溯了这一行为的数值来源并提出了缓解措施 [2]。在我们之前的工作中，即使生成的文本没有变化，我们也展示了词元概率能够揭示大语言模型非确定性 [4]。这些工作涵盖了文本模型，但没有测量图像生成。

B. 确定性推理方案

有多种方案可以实现大语言模型推理的可复现性。批次不变内核会约束每个运算符的归约顺序，使结果不受批次形状影响；Thinking Machines [3]介绍了这种方法，vLLM [6] 和 SGLang [5]则实现了该方法。RepDL 提供逐位可复现的训练和推理 [10]。LLM-42 支持快速但不可复现的路径，并验证结果；如果结果不符合规范，就会回滚 [9]。CoRun 会隔离每个请求的预填充阶段，并采用固定形状的批量解码[7], ；BlockLock 则会固定受保护请求的张量几何结构，使其数值计算路径不受其他请求的影响 [8]。GeulPU [20] 是一款开源引擎，旨在为 GGUF 大语言模型提供逐位相同的 logit。上述方案都针对文本模型。对于扩散模型，SynerDiff 等服务领域的工作研究的是如何通过连续批处理提升速度，而非可复现性 [21]。据我们所知，尚未有针对文本到图像模型的同类解决方案经过评估；我们的结果衡量了此类解决方案需要解决的问题。

C. 文生图模型与量化

文生图生成主要由两大类模型主导。扩散模型根据提示词对潜变量进行迭代去噪，其架构从 SDXL 中的 U-Net [22], 发展到 SD3[23], 、 Qwen-Image [24], 、 HiDream-I1 和 HiDream-O1-Image[25], [26], 以及 FLUX 3 Klein [27]中的多模态扩散 Transformer。自回归模型将图像转换为词元，并依次预测这些词元，例如 Janus-Pro [28], 和 Emu35 [29], 以及 Infinity [30]。在这两类模型中，大规模服务都依赖低精度。扩散模型的量化是一个活跃的研究领域，采用了 Q-Diffusion [31] 等训练后量化方案以及低位宽浮点格式 [32]。本文评估了这些执行选择与可复现性之间的相互作用，而量化文献并未对此进行测量。

D. 图像生成中的非确定性

与我们最接近的研究是 Kim 等人 [11] 和 Fang 等人 [12]。Kim 等人证明，在 Stable Diffusion 中，相同的提示词和随机种子在不同 GPU 上无法生成逐位相同的图像，并据此设计了共识和验证机制。他们的测量针对跨 GPU 和单一模型，并未将批次大小或精度作为影响因素单独考察。Fang 等人在受控因素下评估了扩散语言模型中的非确定性，并表明数据集级指标可能会掩盖这种现象。他们的模型生成的是文本，而非图像。主要推理工具中的从业者报告称，输出会随批次大小和 GPU 而变化，但这些报告缺乏受控测量（参见第 I 节脚注）。这些研究均未针对当前的文生图模型系列，测量由批次大小、量化和硬件引起的非确定性，而这正是本文的研究重点。

III. 方法

A. 模型与硬件

我们评估了涵盖这两个系列的六种文生图模型：扩散模型包括 SDXL base 1.0 [22], 、通义千问-Image-2.1 [24], 、FLUX 3 Klein [27], 和 HiDream-O1 [26] ；自回归生成模型包括 Emu3.5[29] 和 Janus-Pro [28] 。模型运行于来自三个世代的三种 GPU 类型上：NVIDIA RTX A6000（Ampere，不支持 fp8）、RTX 6000 Ada（Ada Lovelace，支持 fp8）和 H100 NVL（Hopper，支持 fp8）。H100 NVL 既以完整设备运行，也以 MIG 3g.47gb 分区运行。表 I 列出了模型和解码配置。主矩阵中的量化变体使用 torchao：int8 为仅权重量化（ HiDream-O1 的语言模型线性层使用 Int8WeightOnly），fp8 则使用 torchao float8 格式。FLUX 3 Klein 采用

批次梯度配置为 8、16 和 32。运行所需的 GPU 时间和峰值内存将在附录（表 **??**）中报告。

B. 提示词

提示词按 S 组组织，每组包含 P 个提示词；每组使用一个基础随机种子，各组之间使用不同的随机种子。SDXL、FLUX 3 Klein 和 HiDream-O1 使用 $S = 3$ 组，通义千问-Image-2.1 使用 $S = 2$。各阶段采用嵌套设置：每组 $P = 10$、25 和 100 个提示词。自回归模型在主要对比中使用$P = 10$ 和 $S = 1$，Emu3.5 和 Janus-Pro 使用相同的提示词。两者也都运行了每组 100 个提示词的设置（Emu3.5 在 H100 NVL 上使用 bf16 和 int8；Janus-Pro-7B 在 RTX 6000 Ada 和 RTX A6000 上使用 bf16）。提示词涵盖多个类别：高频周期图案（棋盘格、光栅和倾斜条纹）、几何和有机形态场景、彩色图案，以及包含数字、文本和计数器的提示词，例如记分牌和七段显示器。这些图案接近模型的分辨率极限，细微的数值变化最容易在此显现。含文本的提示词用于探测语义内容是否正确。图1展示了六个示例提示词对应的各模型批处理大小为 1 时的参考图像。

C. 运行协议

对于每个模型、精度和提示词，我们分别单独生成图像、在批处理中生成图像，并进行重放检查。单独实验组 A0 和 A1 均以批次大小 1 生成每个提示词的图像，A0 (r0) 则是所有比较的参考。批处理分支 B 在每批 8 个提示词中生成图像，提示词成员轮换，位置顺序随机打乱。重放组会原样重复一次运行，以检验不同运行之间的确定性：SDXL、Qwen-Image-2.1 和 HiDream-O1 使用实验组 C 重复第 0 轮的批处理；FLUX 3 Klein、Emu3.5 和 Janus-Pro-7B 使用实验组 R 重放实验组 B（对于 FLUX 3 Klein，重放的是带有 $n = 64$ 的阶段 10 中的 B）。自回归模型采用贪心 argmax 解码，因此不使用采样种子（Janus-Pro 还运行了一个采样实验组，$T = 1.0$，cfg 5.0）。它们的批处理分支 B 使用一个包含 8 个提示词的单元和一个包含 2 个提示词的单元。Emu3.5 在阶段 10 中，实验组 A0、A1、B 和 R 各运行 10 张图像；在阶段 100 中，每组运行 100 张图像。Janus-Pro-7B 在两种解码模式下均运行相同的四个实验组：阶段 10 中每次运行 80 张图像，每种模式 40 张；阶段 100 中每次运行 800 张图像，对应 100 个提示词。对于扩散模型，每个单元均使用相同配置重复 2 至 3 次。表 II 给出了各模型的协议规模。协议由 P、S、随机种子和批处理单元参数化，因此扩大运行规模属于配置变更，而非新协议。完整研究分为三个嵌套阶段，每组分别包含 $P = 10$、25 和 100 个提示词，每组使用一个随机种子。每个阶段都会扩展前一阶段，因此各阶段的结果仍是最终运行结果的有效子集。对于 SDXL，阶段100的运行涵盖一组包含 100 个提示词的提示词（该阶段使用 $S = 1$）。附录 **??** 介绍运行时间和内存使用情况。

D. 指标

我们报告每张批处理图像与其批次 1 参考图像之间的四项指标：位级一致性、平均绝对误差、结构相似性指数和学习感知图像块相似度（表 III）。对于逐位相同的图像对，这些指标均精确为零，因此无需运行网络即可报告。比较前，所有图像均解码为 8 位 RGB；比较过程为只读且离线进行。Qwen-Image-2.1 的 PNG 图像带有阿尔法通道，在计算结构相似性指数、平均绝对误差和学习感知图像块相似度时会忽略该通道。SDXL 的学习感知图像块相似度值以及 Janus-Pro-7B 的跨 GPU 学习感知图像块相似度值，均使用独立的学习感知图像块相似度软件包在 CPU 上计算（AlexNet 骨干网络、线性校准）；这套分析流程不同于图形处理器端的流水线，但两者的学习感知图像块相似度权重逐字节相同。

IV. 结果

本节介绍实证结果。原始生成图像（A0、A1、B 和 C/R 组）以及一个用于检查匹配图像对的交互式查看器已发布在 Zenodo[2] 上。

A. 回放控制

回放。在此报告的各项回放控制中，对于所测试的模型和精度，重复完全相同的运行（使用相同的提示词、随机种子、批次组成和进程设置）都会产生逐位相同的输出。因此，下面报告的差异并非随机噪声，而是由执行设置导致的。举例来说，HiDream-O1 在第 3 组中的三次 fp8 运行结果完全一致。重复运行 Emu3.5 int8 也与之前的运行精确匹配：单独处理和批处理的输出在回放时均逐位相同，各自涵盖 10 个提示词。表 IV 总结了回放控制。在最大规模下，控制结果同样成立。对于 SDXL，批次1 的重复运行在两个阶段运行中均位精确匹配，并且每次运行中的所有 16 张回放图像都与单独处理的参考图像不同（16 张中 16 张）。在阶段100，两个单独重复运行对全部 100 个提示词均位精确匹配（200 对中 200 对），两个批次重放也都与批处理结果完全一致（16 张中 16 张），同时均与单独处理的结果不同（16 张中 16 张）。Emu3.5 在阶段100也表现相同：H100 NVL 上的一次 bf16 运行每个实验组生成 100 张图像（A0、A1、B、R），所有 A0/A1 和 B/R 图像对均与 PNG 精确匹配，全部 400 次解码控制重放均达到像素级精确匹配，int8 阶段100运行的回放结果也完全一致。尽管每次回放都精确匹配，阶段100 的批次与单独运行之间仍存在较大差异。按 bf16 和 int8 的顺序，平均学习感知图像块相似度为 0.296 和 0.307（最大值为 0.673 和 0.804），平均结构相似性指数为 0.598 和 0.597（最小值为 0.008 和 0.013，最大值为 0.987 和 0.994），平均绝对误差为 34.571 和 32.849（最小值为 2.028 和 0.662，最大值为 115.465 和 113.596）；两种精度下，精确匹配的图像对均为 0 对（100 对中）。

B. 批量与单图

概述。表 V 总结了各模型与精度下批处理与单图生成的比较。该表保留了已测量的配置；没有已测量指标的单元仍标记为未测量。新的图形处理器特定运行单独报告，而不是替换到另一图形处理器行中。

[2]https://doi.org/10.5281/zenodo.24444740 3

SDXL fp16 和 bf16 的数据来自数据类型探测：8 个提示词，比较批处理大小为 8 与批处理大小为 1 的情况。

表 I 模型与解码配置。

模型	系列	参数	精度	分辨率	步数	引导强度
SDXL base 1.0	扩散（U-Net）	~3.5B	fp16，bf16	1024×1024 512×512	30	7.5
Qwen-Image-2.1	扩散（MMDiT）	7B	bf16、fp8	768×768	40	1.0（关闭）
FLUX 3 Klein	扩散（整流流）	5.6B	bf16, int8, fp8	512×512	50	3.5
HiDream-O1	扩散	8B	bf16, int8, fp8	512×512	50	5.0
Emu3.5	自回归	34B	bf16，int8	512×512 （32×32 词元）	–	贪心
Janus-Pro-7B	自回归	7B	bf16	384×384 （576 个词元）	–	贪心，采样

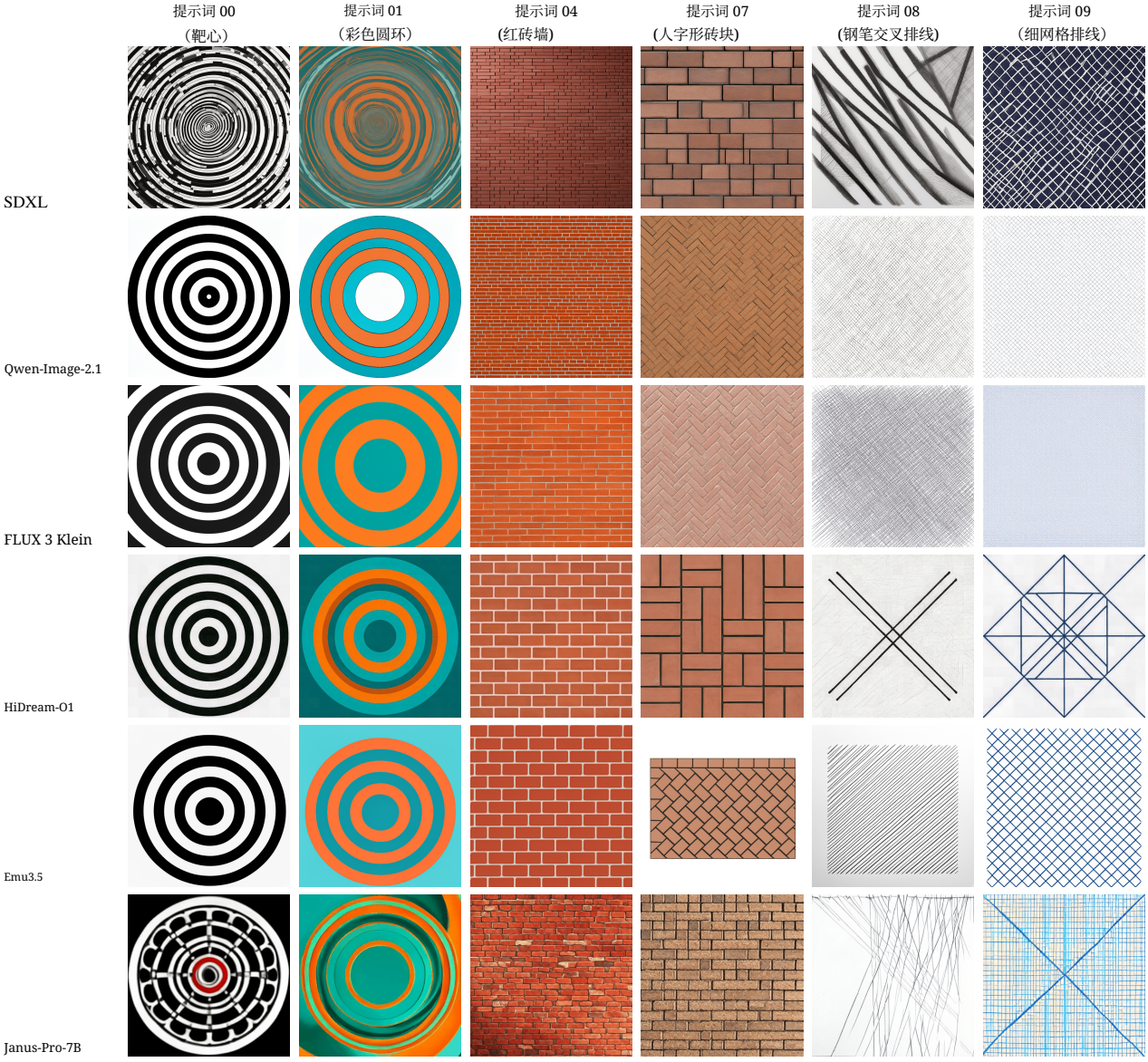


图1. 六个示例提示词（列：提示词00：靶心、提示词01：彩色圆环、提示词04：红砖墙、提示词07：人字形砖块、提示词08：铅笔交叉排线、提示词09：细网格排线）对应的各模型单次运行的批次 1 参考图像（A 臂，首次运行），这些图像未经基于指标的筛选。SDXL 以 fp16、1024 px 运行，Janus-Pro-7B 以原生 384 px 运行，其他模型以 512 px 运行。Janus-Pro-7B 的提示词均以“A detailed image of ...”的形式包裹。

Qwen-Image fp16 行使用 H100 NVL 图像，时间为 512×512，并与 768 像素所选图像对研究分开报告。192 对中的每一对都比较同一阶段中批处理大小为 8 的图像与相同提示词对应的单独生成图像：P10、P25 和 P100 各 64 对，权重相同。不同阶段的提示词有重叠。指标基于丢弃 PNG alpha 通道且不进行合成后的 RGB 计算。精确匹配表示解码后的 RGB 逐位相同；192 对中没有一对符合精确匹配。SSIM 和 LPIPS 均在 CPU 上独立重新计算。

关键观察。有三点观察贯穿整张表。

1) 批次1重跑的结果位精确，而批处理图像在几乎所有情况下都有所不同：决定输出的是批处理，而非随机种子。2) 这种影响的大小取决于模型和精度，范围从位不变（bf16 和 int8 下的 HiDream-O1）到结构性变化（bf16 和 int8 下的 Janus-Pro-7B、Emu3.5，fp8 下的 HiDream-O1，以及 bf16 下的通义千问-图像）。

表 II 各模型和阶段的协议规模：每组提示词数（P）、组数（S）和每个实验组的图像数。A0/A1：两个单独实验组；B：批处理；C 或 R：重放（SDXL、Qwen-Image-2.1 和 HiDream-O1 使用 C；FLUX 3 Klein、Emu3.5 和 Janus-Pro-7B 使用 R）。Qwen-Image-2.1 的 B 组在每个阶段均为 64（8 轮，每轮 8 个）。Janus-Pro-7B 的规模按解码模式划分。每次运行的总数（生成的 PNG 图像）见附录**??**。

模型	Run	P	S A0/A1	B	C/R
SDXL	s10/s25	10/25	3 10/10	25/25 256/632	16
SDXL（bf16）	s100	100	1 100/100	2504	16
Qwen-Image-2.1	s10	10	2 10/10	64	8
Qwen-Image-2.1	s25	25	2 25/25	64	8
Qwen-Image-2.1	s100	100	2 100/100	64	8
Qwen-Image-2.1	s100, B24	100	2 100/100	192	24
FLUX 3 Klein	s10	10	3 10/10	64	64
FLUX 3 Klein	s25	25	3 25/25	96	96
FLUX 3 Klein	s100	100	3 100/100	344	344
HiDream-O1	s10	10	3 10/10	128	8
HiDream-O1	s25	25	3 25/25	128	8
HiDream-O1	s100, B8-B32	100	3 100/100	128/256/384/512	8/16/24/32
Emu3.5	s10	10	1 10/10	10	10
Emu3.5（bf16）s100		100	1 100/100	100	100
Emu3.5 (int8) s100		100	1 100/100	100	100
Janus-Pro-7B s10		10	1 10/10	10	10
Janus-Pro-7B (bf16) P100		100	1 100/100	100	100

表 III 每张批处理图像与其批次 1 参考图像之间的四项指标。

指标	衡量内容
位级一致性	逐位相同图像所占的比例，其中逐位相同是指解码后的像素数组完全相等
MAE	像素的平均绝对误差，取值范围为 0 至 255
SSIM	结构相似性指数 [13]，，使用 scikit-image 0.26.0 在原始分辨率下对 RGB图像计算，数据范围为 255，并对通道取平均
学习感知图像块相似度	感知距离 [14]，使用 AlexNet 骨干网络，使用 lpips 0.1.4，输入缩放至 $[-1, 1]$

表 IV 回放控制：对同一次运行进行相同的重跑。绿色：重跑与原始运行位精确一致。浅灰色模型单元表示自回归模型（Emu3.5、Janus-Pro-7B）。

模型	精度	批次1重跑	批次回放
SDXL fp16，bf16[3]		yes	yes
Qwen-Image-2.1	bf16、fp8	yes	yes
FLUX 3 Klein	bf16、int8、fp8	yes	yes
HiDream-O1	bf16、int8	yes	yes
HiDream-O1	fp8	是（运行 3 次）	yes
Emu3.5	bf16，int8	yes	yes
Janus-Pro-7B	bf16	yes	yes

3) 在大多数情况下，批处理图像虽然不同，但同样有效；例外情况将在第四节-C中讨论。

SDXL，采用 fp16 精度。采用 fp16 精度的 SDXL 中，每张批处理图像都会发生变化，但变化幅度很小。运行器记录的 B 组平均绝对误差为：第 10 阶段的 256 对图像中为 1.85（最差的一对：提示词 08，平均绝对误差 4.98）；第 25 阶段的 632 对图像中为 2.164（最差的一对：提示词 23，平均绝对误差 8.72）。两个阶段中，B 组都没有图像与另一张逐位相同。在 B 组内，第 10 阶段的 3150 对相同提示词图像中有 197 对出现差异，第 25 阶段的 7675 对中则有 0 对出现差异。本地学习感知图像块相似度分析显示，B 组在两个阶段的均值均为 0.013（$n = 256$ 和 $n = 632$）；C 组与批次 1 参考图像相比，均值分别为 0.011 和 0.007（各为 $n = 16$）；单独组重复运行的均值则恰好为 0（$n = 20$ 和 $n = 50$）：单独重跑的结果位精确。B 组最差图像对的学习感知图像块相似度分别达到 0.045（第 10 阶段，提示词 06）和 0.097（第 25 阶段，提示词 23）。图 2 总结了第.....阶段对精度的依赖

表 V 批量与单图生成：使用相同的提示词和随机种子。与批次 1 参考图像相比的结构相似性指数（1.0 表示相同）。阴影标示：绿色 = 相同或近乎相同，黄色 = 低于可见阈值或仅有轻微可见差异，橙色 = 可见的结构变化，红色 = 显著的结构变化，深灰色 = 未测量。浅灰色无结果单元格标示自回归模型（Emu3.5、Janus-Pro-7B）。SDXL 的平均绝对误差由运行器记录。

模型	精度	图形处理器	逐位标识	平均绝对误差	结构相似性指数	学习感知图像块相似度
SDXL	fp16	A6000	0.0	3.308		
SDXL	bf16	A6000	0.0	8.230		
SDXL (s10)	fp16	A6000	0.0	1.85 0		
SDXL (s25)	fp16	A6000	0.0	2.164		
Qwen-Image-2.1	fp16	H100 NVL	0.000	6.789		
Qwen-Image-2.1	bf16	H100	0.0	7.476		
Qwen-Image-2.1（s10）	bf16	H100	0.0	5.54	0.894	
Qwen-Image-2.1（s10）	fp8	H100	0.0	7.75	0.811	
Qwen-Image-2.1 (s25)	bf16	H100	0.0	7.98	0.850	
Qwen-Image-2.1 (s25)	fp8	H100	0.0	9.72	0.820	
Qwen-Image-2.1（s100）	bf16	H100	0.0	8.91 0		
Qwen-Image-2.1 (s100)	fp8	H100	0.0	10.00		
FLUX 3Klein bf16		H100-M	0.0	0.173		
FLUX 3Klein(s10) bf16		H100-M	0.0	0.359		
FLUX 3 Klein (s10)	int8	H100-M	0.0	0.356		
FLUX 3 Klein (s10)	fp8	H100-M	0.0	17.33	0.829	0.074
FLUX 3 Klein (s100)	fp8	H100-M	0.0	8.86	0.893	
FLUX 3 Klein (s100)	bf16	H100-M	0.0	0.127		
FLUX 3 Klein (s100)	int8	H100-M	0.0	0.117		
HiDream-O1	bf16	H100-M	1.00	0.0 1.0		
HiDream-O1	int8	H100-M	1.00	0.0	1.000	
HiDream-O1	fp8	H100-M	0.0	15.45	0.748	0.109
HiDream-O1（s25）	fp8	H100-M	0.0	17.85	0.731	0.159
HiDream-O1 (s100)	bf16	H100-M	1.00	0.0 1.0		
HiDream-O1 (s100)	int8	H100-M	1.00	0.0	1.000	
HiDream-O1（s100）	fp8	H100-M	0.0	19.56	0.686	0.176
Emu3.5	bf16	H	0.0	30.16	0.580	0.168
Emu3.5	int8	H100	0.0	31.78	0.551	0.229
Janus-Pro-7B（采样）	bf16	/	0.0	46.64	0.261	0.296
Janus-Pro-7B（贪心）	bf16	/	0.0	54.77	0.229	0.319

10，六个模型共有的唯一阶段。

SDXL 在第 100 阶段使用 bf16。在第 100 阶段使用 bf16（包含一组 100 个提示词，共 313 轮、每轮批处理 8 个），2504 张批处理图像全部不同于批次 1 参考图像（2504 张中 0 张精确匹配），平均学习感知图像块相似度为 0.072，最大值为 0.359（提示词 46）。对于相同的 2504 对批处理图像与其自身提示词 A0 图像，平均结构相似性指数为 0.876（最小值 0.464，最大值 0.982），平均绝对误差为 8.854（最小值 1.755，最大值 42.133）。结构相似性指数采用原生 1024 像素 RGB 数组，数据范围为 255，并进行通道平均；平均绝对误差使用 0–255 范围内的所有 RGB 通道。汇总均值对每一对赋予相同权重：96 个提示词各有 25 对，4 个提示词各有 26 对。

RTX 6000 Ada 上的 Qwen-Image。在 512×512 运行的另一个 bf16 阶段 100 中，共有 272 张图像，且没有图像失败标记。所有 100 次单独重复运行均与各自的单独参考图像相符；所有 64 张批处理图像和 8 张重放图像均与单独运行的结果不同，而所有 8 张重放图像均与对应的第 0 轮批处理结果相符。该阶段没有批内重复配对，因此本次运行无法检验伴随不变性。本次运行耗时 2236.5 秒，预留内存峰值为 48,932,847,616 字节，采样进程内存为 47,180 MiB。由于 GPU/运行时不同，这些计数与 H100 指标行分开报告。对 64 对批处理与单独运行结果进行 CPU 评分，在舍弃 alpha 通道且不进行合成后，RGB 图像的汇总结构相似性指数（SSIM）为 0.894，学习感知图像块相似度（LPIPS）为 0.070。这些指标

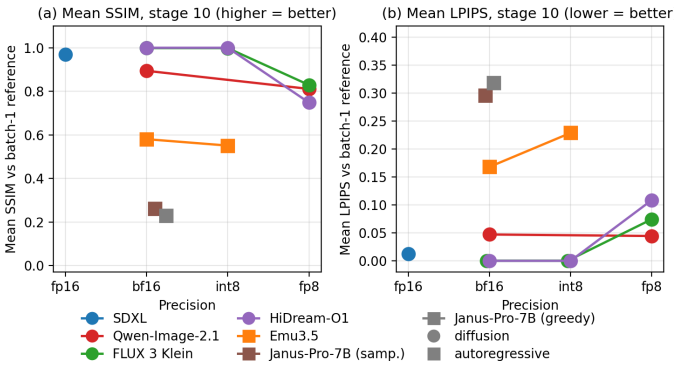


图 2. 精度对第 10 阶段的影响（这是六个模型共有的唯一阶段）：批处理图像与批次 1 参考图像相比的平均结构相似性指数 (a) 和平均学习感知图像块相似度 (b)。圆形：扩散模型；方形：自回归模型。六个模型均显示学习感知图像块相似度。

属于这次历史性的 512像素运行，而非另行进行的固定输入 768 像素对比。

512 像素下的 Qwen-Image。在 512×512，Qwen-Image 在两种精度下都呈现出明显的批次效应，重放组也证实了这一效应具有确定性：B 组在每个阶段都生成 $n = 64$ 张图像，C 组以 $n = 8$ 张逐位相同的图像重放第 0 轮批处理，而 B 组在任何阶段都没有图像逐位相同。结构相似性指数是在丢弃阿尔法通道后基于 RGB 计算的。按 bf16/fp8 顺序，B 组的平均值在第 10 阶段为 0.894/0.811（逐提示词为 0.898/0.817，最小值为 0.667/0.135），第 25 阶段为 0.850/0.820（逐提示词为 0.850/0.828，最小值为 0.104/0.131），第 100 阶段为 0.875/0.863（最小值为 0.268/0.520）。C 组的平均值在第 10 阶段为 0.901/0.826（最小值为 0.667/0.298），第 25 阶段为 0.877/0.882（最小值为 0.661/0.810），第 100 阶段为 0.924/0.951（最小值为 0.716/0.920）。

FLUX 3 Klein。FLUX 3 Klein 仅包含组 A 和组 B，在 bf16 和 int8 下的表现几乎完全相同：第 10 阶段（B $n = 64$）两者的平均结构相似性指数均约为 0.998，第 25 阶段（$n = 96$）均约为 0.999；两个阶段中，两者的最低值均约为 0.989。只有 fp8 打破了这一规律：第 10 阶段的平均结构相似性指数降至 0.829（逐提示词均值为 0.836，最低值为 0.260），第 25 阶段降至 0.855（逐提示词均值为 0.865，最低值为 0.025），且两个阶段都没有图像逐位相同。

HiDream-O1。HiDream-O1 是唯一一个在批处理下保持逐位不变的模型：在 bf16 和 int8 下，每张批处理图像在所有阶段都与其批次 1 参考图像完全相同（第 3 组的全部 128 张图像均逐位相同，结构相似性指数为 1.000；阶段 25 的 161 对已评分学习感知图像块相似度均精确匹配；阶段 100 的 A 组（$n = 100$）、B 组（$n = 128$）和 C 组（$n = 8$）也均逐位相同，结构相似性指数为 1.000、平均绝对误差为 0、学习感知图像块相似度为 0）。只有 fp8 存在差异，而且差异很大且稳定：三次运行都能完全复现。在第 3 组中，平均结构相似性指数为 0.725（最低值为 0.302），每张图像都不相同，最差提示词对应的数值分别为 0.68 和 0.69。按阶段来看：第 10 阶段为 0.748（最低值为 0.554），平均绝对误差为 15.45，学习感知图像块相似度为 0.109（最高值为 0.320）；阶段 25 为 0.731（B 组 $n = 128$ 的最低值为 0.296），平均绝对误差为 17.85，学习感知图像块相似度为 0.159（最高值为 0.652）；其中 C 组（$n = 8$）为 0.731（最低值为 0.646），平均绝对误差为 15.95，学习感知图像块相似度为 0.128（最高值为 0.259），提示词 19 在两个指标上都是最差的；阶段 100 为 0.686（最低值为 0.255），平均绝对误差为 19.56，学习感知图像块相似度为 0.176（最高值为 0.636）；其中 C 组为 0.697（最低值为 0.318），平均绝对误差为 24.40，学习感知图像块相似度为 0.235（最高值为 0.590），提示词 41 再次表现最差

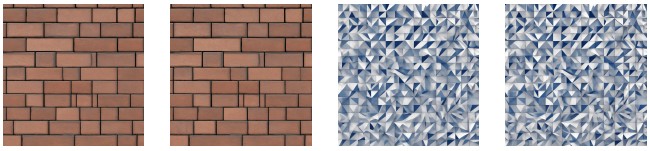


图 3。SDXL，fp16，RTX A6000，批处理与批次 1 的对比。结构相似性指数最低的一对（阶段 10，提示词 07，砖墙）：结构相似性指数为 0.877；学习感知图像块相似度最低的一对（阶段 25，提示词 23）：学习感知图像块相似度为 0.097。每对图像中，左侧为批次 1 参考图像；右侧为批处理图像。

两者。

自回归模型。自回归模型呈现出最大的总体效应。Emu3.5 在 bf16 下的平均结构相似性指数为 0.580，在 int8 下为 0.551；全部 10 个提示词与批次 1 参考图像均不同（最低值分别为 0.216（平均绝对误差为 40.07）和 0.308；平均学习感知图像块相似度分别为 0.168 和 0.229，最高值分别为 0.496 和 0.472）。在 RTX 6000 Ada 上运行的 Janus-Pro-7B（bf16）采用采样解码时（$T = 1.0$，cfg 5.0）平均结构相似性指数为 0.261，采用贪心解码时为 0.229（最低值分别为 0.127 和 0.080；学习感知图像块相似度分别为 0.296 和 0.319，最高值分别为 0.474 和 0.595）。重放组能精确复现批处理图像，因此，即使采用采样，效应也是确定性的，而非噪声。该效应在 100 个提示词下仍然存在。在 RTX 6000 Ada 上，批处理与单独处理的学习感知图像块相似度均值分别为采样解码 0.368、贪心解码 0.367；在 RTX A6000 上分别为 0.381 和 0.347（最高值介于 0.645 至 0.852）。四种组合中，100 对图像均无一精确匹配；每次单独重复运行和批次回放都能精确复现。RTX 6000 Ada 上，采样解码和贪心解码的平均结构相似性指数分别为 0.400 和 0.418（最低值分别为 -0.010 和 0.050；平均绝对误差均值分别为 42.704 和 42.892，最高值分别为 141.809 和 104.883）；RTX A6000 上，采样解码和贪心解码的平均结构相似性指数分别为 0.409 和 0.435（最低值分别为 0.069 和 0.067；平均绝对误差均值分别为 41.018 和 40.037，最高值分别为 100.840 和 101.873）。

C. 最差案例

扩散效应最大。我们测得最大的扩散效应出现在 Qwen-Image 使用 bf16、分辨率为 768 像素时：对于包含 75 度红奶油色条纹的提示词，批次大小为 8 时，条纹会变成人字形交叉网格（结构相似性指数 0.228，学习感知图像块相似度 0.388）；而使用 fp16 处理相同提示词时，条纹仍然清晰，结构相似性指数为 0.997（图 4）。

数字和文本。表 VI 汇总了测得的最差案例。图 3 至图 8 按模型展示了差异最大的配置中，结构相似性指数最差和学习感知图像块相似度最差的图像对所对应的参考图像与批处理图像。在大多数最差案例中，批处理图像仍然看起来合理。例外情况是包含数字、文本和计数器的提示词，此时批处理图像明显较差：通义千问图像模型使用 bf16 生成的记分牌中，批处理图像的数字错误或损坏，而批次 1 参考图像则是正确的（图 4）。对于保留的 Emu3.5 bf16 七段显示器图像对（提示词 102），两张图像都未能正确呈现要求的 88:88。批处理会改变这对图像，但这并不能证明原本正确的单独生成图像会因此变得不正确。

较早设置的 HiDream-O1 fp8 行报告的是提示词均值的非加权平均（每组 24 个提示词和 128 个列出的成对项），而非最差成对项；汇总后的学习感知图像块相似度分别为 0.180 和 0.171。每组中除最差项外的 125 个成对项，是根据逐对行核查的，而不是从像素重新评分得出的。通义千问 Qwen-Image fp16 的学习感知图像块相似度值为正，但四舍五入后为 0.000（小于 0.001）。通义千问 Qwen-Image 的成对指标会丢弃 alpha 通道，不进行合成；

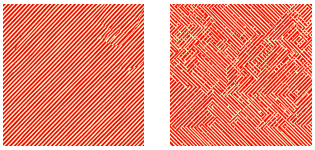


图 4。Qwen-Image-2.1，bf16，768 像素，H100 NVL，批次 8 与批次 1 对比。最差的一对（提示词 11，75 度红奶油色条纹）：结构相似性指数 0.228，学习感知图像块相似度 0.388。左：批次 1 参考图像；右：批处理图像。同一对在结构相似性指数和学习感知图像块相似度上均为最差。

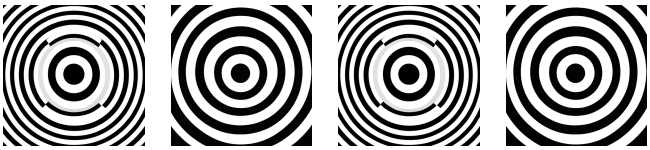


图 5。FLUX 3 Klein，fp8，H100 NVL，批处理与批次 1 对比。最差的结构相似性指数对（提示词 00，同心圆环）：结构相似性指数 0.260，学习感知图像块相似度 0.343，平均绝对误差 102.5；最差的学习感知图像块相似度对（提示词 00）：学习感知图像块相似度 0.343。



图 6。HiDream-O1，fp8，H100 NVL，批处理与批次 1 对比。最差的结构相似性指数对（第 3 组，提示词 19，倾斜条纹）：结构相似性指数 0.302，平均绝对误差 68.97；展示的最差学习感知图像块相似度对（第 3 组，提示词 19）：学习感知图像块相似度 0.652。选定的结构相似性指数对的对应学习感知图像块相似度为 0.644；0.652 这一对是归档的学习感知图像块相似度最高值。

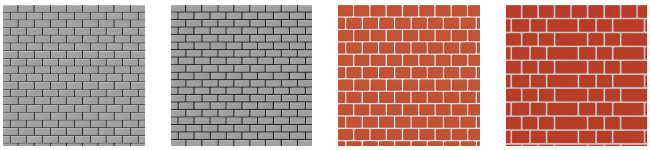


图 7。Emu3.5，int8，H100 NVL，批处理与批处理 1 对比，阶段10。最差的结构相似性指数配对（提示词 5，灰色砖墙）：结构相似性指数 0.308；最差的学习感知图像块相似度配对（提示词 4）：学习感知图像块相似度 0.472。

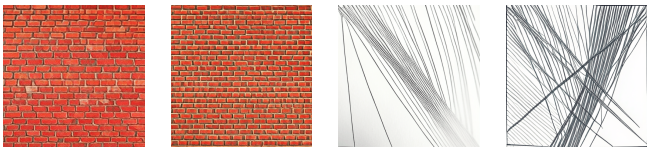


图 8。Janus-Pro-7B，bf16，RTX 6000 Ada，批处理与批处理 1 对比，贪心解码。最差的结构相似性指数配对（提示词 4，砖墙）：结构相似性指数 0.080，学习感知图像块相似度 0.390；最差的学习感知图像块相似度配对（提示词 8，线条图案）：学习感知图像块相似度 0.595，结构相似性指数 0.122。

记分牌上的结构相似性指数在白色背景下变为 0.514，在黑色背景下变为 0.328。对于原始的 Emu3.5 p102 单独生成与 B8 的成对图像，结构相似性指数和平均绝对误差分别经过独立重新计算。对保留的这组成对图像进行 CPU 评分，得到学习感知图像块相似度 0.470。两张图像都未能正确呈现所要求的 88:88。

对于 HiDream-O1 fp8，使用结构相似性指数和学习感知图像块相似度重新计算了组数 1、2 和 3 中选出的三个结构相似性指数最差配对。组数 3 的配对为 p19，第 09 轮，槽位 1。这些是针对特定配对的分数，而不是某组内学习感知图像块相似度的最大值。每组其余 127 个批处理配对（总计 381 个）保留已存档的结构相似性指数值，本次检查未重新评分。较早组数的提示词均值摘要另行保留。

表 V 中 SDXL bf16 数据类型探测的学习感知图像块相似度单元使用了来自八个提示词的 16 组批处理与单独处理配对，每次批处理调用都以相同方式重复；它并非阶段100样本池。对于表 VII 中 fp16 阶段100 B8 条目，结构相似性指数基于 2504 组配对和 100 个提示词汇总，其中逐位相同的配对数为零，且

表 VI 测得的最差情况。阴影标示同表 V。

Case	模型	精度	结构相似性指数	学习感知图像块相似度
交叉影线，768 像素，B8	Qwen-Image-2.1	bf16	0.228	0.388
相同提示词	Qwen-Image-2.1	fp16	0.997	≈0
选定的结构相似性指数配对，第 3 组（p19）	HiDream-O1	fp8	0.302	0.644
			0.679	0.180
较早的组（均值）	HiDream-O1	fp8	0.687	0.169
第 1 组，p19，r01，槽位 0	HiDream-O1	fp8	0.308	0.535
第 2 组，p10，r13，槽位 6	HiDream-O1	fp8	0.294	0.485
最差提示词（p5）	Emu3.5	int8	0.308	0.289
最差提示词（p5）	Emu3.5	bf16	0.216	0.292
最差提示词，s10	HiDream-O1	fp8	0.554	0.320
最差提示词，s10	SDXL	fp16	0.877	0.045
最差提示词，s10	FLUX 3 Klein	fp8	0.260	0.343
最差提示词，s10	Qwen-Image-2.1	bf16	0.667	0.289
最差提示词，s10	Qwen-Image-2.1	fp8	0.135	0.083
最差提示词，s25（p23）	SDXL	fp16	0.875	0.097
最差提示词，s25	FLUX 3 Klein	fp8	0.025	0.584
最差提示词，s25	Qwen-Image-2.1	bf16	0.104	0.122
最差提示词，s25	Qwen-Image-2.1	fp8	0.131	0.083
最差提示词，s25（p19）	HiDream-O1	fp8	0.296	0.652
最差提示词，s100	FLUX 3 Klein	fp8	0.274	0.343
最差提示词，s100	FLUX 3 Klein	bf16	0.989	0.002
最差提示词，s100	FLUX 3 Klein	int8	0.989	0.002
最差提示词，s100	Qwen-Image-2.1	bf16	0.268	0.233
最差提示词，s100	Qwen-Image-2.1	fp8	0.520	0.388
最差提示词，s100 (p41)	HiDream-O1	fp8	0.255	0.636
最差提示词，采样	Janus-Pro-7B	bf16	0.127	0.474
最差提示词，贪心	Janus-Pro-7B	bf16	0.080	0.595
记分牌数字	Qwen-Image-2.1	bf16	0.499	0.626
七段数字 (p102)	Emu3.5	bf16	0.254	0.470

范围为 0.875–0.997。由于批次输出在各轮中完全重复，因此仅出现 100 个不同的配对得分；所有低于 0.92 的得分都来自提示词 7、8、23 和 99。

D. 批次大小

剂量-响应。批次大小的影响远小于批处理本身的影响：对于每个设有批次梯度的模型，从批次1增至批次8造成的下降，远大于从批次8增至32的任何变化。本节仅改变批次大小：模型、精度、提示词集和随机种子均保持不变，评分池则随批次大小增加，具体如下所述。对于采用 bf16 的 FLUX 3 Klein，批次8增至批次32并未使影响加剧，两种情况下 SSIM 均不低于 0.989，LPIPS 均不高于 0.002。通义千问 Qwen-Image 是唯一一个随着批次大小增加、池化 SSIM 持续下降的模型，而且下降幅度平缓：fp8 下，池化 SSIM 从批次8时的 0.863 降至批次16时的 0.837（最小值 0.197）、批次24时的 0.833（最小值 0.221）和批次32时的 0.827（最小值 0.165）；bf16 下则从批次8时的 0.875 降至批次16时的 0.864（最小值 0.274）和批次24时的 0.857（最小值 0.116）。逐提示词最小值下降得更快：fp8 下从 0.520 降至 0.165，bf16 下从 0.268 降至 0.116。LPIPS 序列并非单调变化，故另行报告。在 512×512 H100 的批次梯度测试中，批次32、bf16 的池化 SSIM 达到 0.853，LPIPS 为 0.089，涵盖来自 100 个提示词的 256 组批处理与单张图像配对。HiDream-O1 在 bf16 和 int8 下，批次8、16、24 和 32 的结果与批次1逐位相同；在 fp8 下，批次16时结果接近 0.68（SSIM 0.676，最小值 0.246，LPIPS 0.181），批次24时为（SSIM 0.683，最小值 0.223，LPIPS 0.176），批次32时为（SSIM 0.678，最小值 0.204，LPIPS 0.186），与批次8时的 0.686 几乎没有变化。对于 SDXL，在 46 GB 十进制容量的

表 VII 批次大小剂量-响应。阴影标示同表 V。FLUX 3 Klein 的批处理分支将批次 8、16 和 32 合并，因此没有单独的 B16 和 B24 测量值；效果未显示批次趋势。未尝试表示没有阶段结果。SDXL 的 B16 以及单独的批次 24 和批次 32 调用在 46 GB 十进制分配器上限下内存不足。

模型	精度	指标	B8	B16	B24	B32
FLUX 3 Klein	bf16	结构相似性指数	≥0.989	汇总于 B8 和 B32		≥0.989
FLUX 3 Klein	bf16		≤0.002	汇总于 B8 和 B32		≤0.002
SDXL	fp16	结构相似性指数	0.977	内存不足		
Qwen-Image-2.1	bf16	结构相似性		0.864		
Qwen-Image-2.1	fp8	结构相似性		0.837		
Qwen-Image-2.1	bf16	学习感知图像		0.093		
Qwen-Image-2.1	fp8	学习感知图像		0.098		
HiDream-O1	bf16	SSIM	1.000	1.000		
HiDream-O1	bf16	学习感知图像		0.0	0.0	0.0
HiDream-O1	int8	SSIM	1.000	1.000		
HiDream-O1	int8	学习感知图像		0.0	0.0	0.0
HiDream-O1	fp8	结构相似性		0.676		
HiDream-O1	fp8	学习感知图像		0.181	0.176	0.186

PyTorch 分配器上限下，RTX A6000 的 VAE 卷积在批次16时发生 CUDA 内存不足。批次24和批次32的独立运行均在完成 32 次单张批次1调用后，分别尝试一次批处理调用。两者均未生成图像，并且在相同上限下都触发了明确的 CUDA 内存不足错误（允许使用 42.84 GiB，物理可用内存为 16.03 GiB；批次32的请求量为 16.00 GiB），因此失败是由上限所致，而非物理内存耗尽。在这张 48 GB 显卡上，批次8仍是测得的最大成功批次；独立运行中最大的成功批次为批次1（32 次全部成功）。对于 Qwen-Image-2.1，批次8、16、24、32 时的池化 SSIM 分别基于 n =64、128、192、256 张批处理图像；对于 HiDream-O1，批次16、24、32 时分别基于 $n = 256$、384、512 张图像，因为评分池会随批次大小增加。Emu3.5 的批次8数值来自 100 个提示词的集合（$n = 100$）。表 VII 给出了剂量-响应数值。这些比较在较大批次下纳入了更多图像，且仅涵盖设有批次梯度的三种模型，因此该趋势是在矩阵子集上测得的。

E. 因素归因

批处理。在扩散模型实验中，固定批处理的输出在所测试的伴随项和位置之间保持稳定。在 RTX 6000 Ada 上进行的一项独立 Janus-Pro-7B bf16 实验测试了四个目标、两组互有重叠的七个伴随项，以及目标位置 0、3 和 7；单独调用和批处理调用的填充长度均固定为 52。对于贪心解码和一个固定的采样随机种子，每个目标的全部六个 B8 目标输出在词元和解码后的像素方面均完全一致：目标内部无序比较共 120/120 对精确匹配。全部 48 个目标 B8 输出都与其单独运行的参考图像不同，而全部 392 对图像级重放结果均精确匹配。这些结果表明，在测试范围内，输出对伴随项和位置保持稳定；但不能据此推及所有自回归模型、提示词或随机种子。

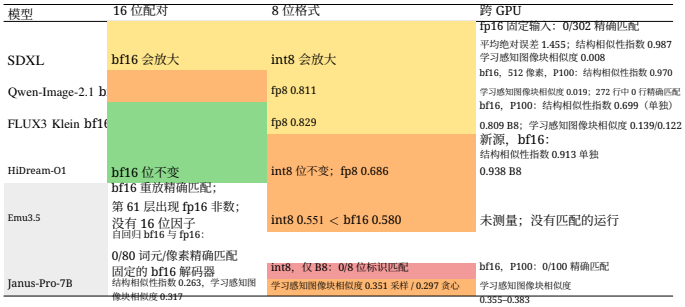
模型家族。架构家族对我们测得的效应大小的影响，远超我们改变的任何其他因素。两个自回归模型呈现出我们测得的最大效应：Emu3.5 在 bf16 下的平均结构相似性指数为 0.580，在 int8 下为 0.551；Janus-Pro-7B 在采样解码下为 0.261，在贪心解码下为 0.229，平均绝对误差最高达 54.77，平均学习感知图像块相似度最高达 0.319；并且四种配置中的每一种都有 10 个提示词在所有四种自回归配置下均与批次 1 参考图像不同；对于两个自回归模型，在 100 个提示词下，批处理与单独运行的 100 对结果中，精确匹配的数量均为 0。没有任何扩散配置接近这一水平：扩散模型的最低平均结构相似性指数为 0.686（HiDream-O1，fp8，第 100 阶段），甚至高于自回归模型的最高平均值；而 bf16 扩散运行的数值均不低于 0.875，其中 HiDream-O1 位不变，FLUX 3 Klein 则不低于 0.989。两者差距显著：自回归模型中最低的平均值（0.229）约为扩散模型中最低平均值（0.686）的三分之一。两类模型中的效应都是确定性的；在各自类别内，精度和模型决定了精确数值，但效应的量级取决于所属类别。第 V 节将讨论这种差异的一种可能机制。

精度。低精度会放大这种效应，但位宽本身并不能决定结果：表现最差的格式取决于模型。观察到的情况仅涉及 8 位格式：FLUX 3 Klein、HiDream-O1 和 Qwen-Image 使用 fp8，Emu3.5 使用 int8。Emu3.5 表现最差的格式是 int8（平均结构相似性指数为 0.551，而 bf16 为 0.580）；FLUX 3 Klein 和 HiDream-O1 则是 fp8（它们的 int8 在第 3 组中位不变，结构相似性指数为 1.000）。对于 Qwen-Image-2.1，bf16 的批处理敏感度在 8 个提示词中的 7 个高于 fp16，而 fp8 的表现比两者都差。对于 SDXL，与 fp16 相比，bf16 和 int8 都会放大这种效应；与此同时，bf16 与参考图像接近（结构相似性指数为 0.96–0.98），且会改变每张图像。在通用 RTX A6000 上对 FLUX 3 Klein 进行的一次 int8 探测（在 Transformer 线性层中使用 torchao Int8WeightOnly，文本编码器和变分自编码器采用 bf16；50 步、28 次调用、148 张图像、34.3 分钟、峰值显存 25.916 GiB）精确复现了其单独重跑和批次回放结果（分别为 0/10 和 0/64 对），但全部 64 对批处理与单独处理结果以及批处理内部 180 对中的 132 对都存在差异；此处的精确匹配指解码后的 uint8 RGB 数组完全相等，而不仅仅是 PNG 文件字节相同。64 对批处理与单独处理结果的平均 LPIPS 为 0.008（最高 0.051）；若 30 对不同的解码结果各计一次，则平均值为 0.007。180 对无序的同提示词批处理内部比较包含重复输出，也包含批次大小为 8、16 和 32 时的交叉比较；其中有 52 对解码结果不同。所有列出的配对的平均 LPIPS 为 0.008，而 30 对不同且存在差异的结果的平均值为 0.010（最高 0.051）。单独重复运行和批次回放的 LPIPS 均为 0.000。这些汇总结果与 H100-M 各行分开报告。

图形处理器。更换图形处理器会改变每一张匹配的 Janus-Pro- 7B bf16 图像（RTX A6000 与 RTX 6000 Ada）。在 100 个提示词上，采样单张图像的跨 GPU 学习感知图像块相似度为 0.383，批处理/重放的结果为 0.368；贪心结果分别为 0.355 和 0.363。GPU内批处理在 Ada 上的学习感知图像块相似度为 0.368/0.367，在 A6000 上为 0.381/0.347（采样/贪心）：这些影响的量级相近。表 VIII 总结了其范围。

在一项独立的 Qwen-Image-2.1 bf16 固定输入对比中，RTX 6000 Ada 与 RTX A6000 之间的跨 GPU 差异小于对 100 多个提示词进行批处理造成的差异。在 272 条组别相同的 RGB 记录中（200 条单独运行、64 条批处理、8 条重放），MAE/SSIM/LPIPS 的均值分别为 2.313/0.970/0.019。这些记录涵盖 100 多个提示词，并非 272 个独立样本：100 条单独运行记录会重复其他单独运行记录。95% 提示聚类自助法区间为 [1.804, 2.857]（MAE）、[0.962,0.978]（SSIM）

表 VIII 各模型的因素归因。阴影标示同表 V。Qwen-Image-2.1 的跨设置比较见正文。



以及 [0.012, 0.026]（LPIPS）。将跨 GPU 记录限制为批处理覆盖的 64 个提示词后，共有 200 条记录，SSIM 为 0.972，LPIPS 为 0.019。以各图形处理器自身的单独运行结果作为参考图像进行批处理时，Ada 上的 SSIM/LPIPS 为 0.898/0.069，A6000 上为 0.907/0.062（各 64 条记录）。按提示词匹配计算的 LPIPS 差值（跨 GPU 减去批处理）分别为 -0.049 [- 0.064、-0.036] 和 -0.042 [-0.057、-0.028], 。这些差值是在每个提示词内先取均值再相减；它们并非混合比例和记录权重不等时汇总均值之差。保留全部 100 个提示词，包括提示词 38。

这一固定输入的跨 GPU 差异测试了使用固定 CPU 潜变量和由 Ada 生成的条件信息的去噪器与解码器，而不是文本编码器。每个 GPU 在 768×768 运行一次，使用 40 个步数、bf16 和随机种子 20261110，并对每张图像进行变分自编码器切片。不排除运行时、cuDNN 和内核选择方面的差异。cuDNN 确定性模式处于关闭状态，允许卷积使用 TF32，而矩阵乘法使用 TF32 处于关闭状态；精确重放并不意味着启用了确定性模式。PNG/原生 RGBA 完整性与 RGB 指标是不同的：评分时会丢弃 alpha 通道，不进行合成。单独重跑中 100/100 个提示词均为精确匹配；每个 GPU 的 8/8 对批次回放也均为精确匹配；本阶段没有重复的 B 对 B 配对。没有任何跨 GPU 行是精确匹配，观测到的精确匹配率约为零；自助法区间[0,0] 并非总体上界。这些结果无法确立图像质量排名。

另一项通义千问 fp8 全尺寸 H100 与多实例 GPU 的比较同时更换了软件和运行环境，因此属于跨设置比较，而非仅比较图形处理器。92 对 RGB 图像中没有一对完全匹配；单独运行的均值为：结构相似性指数 0.949、学习感知图像块相似度 0.018、平均绝对误差 3.888；而 B8means 四舍五入后，结构相似性指数为 1.000，学习感知图像块相似度为 ≈0 （平均绝对误差 0.071）。相似度经四舍五入后看似相同，并不意味着完全相等。

本研究没有在第二块图形处理器上进行匹配的 Emu3.5 运行，因此未测量该模型的跨 GPU 归因；这种缺失既不能证明跨 GPU 效应存在，也不能证明其不存在。

SDXL 固定输入 fp16 s10 运行在每个图形处理器上完成了 64 次生成调用和 302 张图像：30 次 A 重复、32 轮 B 和 2 次 C 重复。这些运行采用 fp16 检查点加载和 UNet 推理，并使用标准的强制升精度变分自编码器解码，而非所有组件统一使用 fp16。记录的实际运行时间在 RTX A6000 上为 2027.426 秒，在 RTX 6000 Ada 上为 2047.642 秒；生成调用时间分别为 1768.860 秒和 1788.380 秒，后处理时间分别为 203.191 秒和 203.401 秒。PyTorch 的峰值为已分配 39.331 GB，A6000 上已预留 39.456 GB，Ada 上已分配 39.331 GB、已预留 39.458 GB。采样得到的 nvidia-smi 峰值在 A6000 上为本进程 39.800 GB（37,956 MiB）、整台设备 39.810 GB（37,966 MiB）；在 Ada 上为本进程 39.999 GB（38,146 MiB）、整台设备 40.009 GB（38,156 MiB）。采样峰值是下限。cuDNN 确定性模式处于关闭状态。

匹配的 A6000 和 Ada 输出在每一对中都存在差异。对 302 对调用、标签和提示词均相同的图像进行离线 CPU 比较后发现，像素精确匹配和文件精确匹配均为 0。总体平均绝对误差、结构相似性指数和学习感知图像块相似度均值分别为 1.455、0.987 和 0.008。各组均值分别为：A 组 1.182/0.991/0.004（30 对）、B 组 1.485/0.987/0.009（256 对），以及 C 组重放 1.491/0.988/0.007（16 对）。各组中重复出现的输出意味着，这 302 行表示重复输出权重，而非 302 个独立的提示词观测值。

Emu3.5 bf16 完整批处理/重放运行完成了 24 次调用并生成了 40 个 PNG。自回归外部墙钟时间为 A0 430.868 秒、A1 428.476 秒、B128.487 秒和 R128.257 秒。全部 20 对重放在词元、解码像素和未归一化 CFG 引导日志分数上均为精确匹配；这些分数不是概率质量、KL 散度或熵。PyTorch 的最大峰值为已分配 79.399 GB、已保留 87.992 GB，采样得到的 nvidia-smi 设备峰值为 87.731 GB（83,667 MiB），这是一个下界。这是 bf16 批处理/重放结果，而非精度比较；由于 fp16 在第 61 层产生非数，因此不可用。

Janus 在固定 bf16 图像解码器的情况下，将自回归 bf16 与 fp16 进行比较，而不是比较端到端 fp16 流水线。列出的 80 个对比包含 10 个提示词上的 40 对唯一配对，汇总了贪心解码和采样解码；80 对中有 0 对为位精确匹配。SDXL 在 8 个提示词上，于每种数据类型内部比较批处理大小 8 与批处理大小 1；其 16 行重复了相同的 8 对结果。这不是 bf16 与 fp16 的比较。SDXL 的数据类型与这两项研究彼此分开处理。各项计数并非独立样本，也未报告置信区间。

在一项独立的新来源 HiDream-O1 bf16 研究中（并非历史供应商配置），匹配的 A6000/Ada 运行在512×512、50 个步数下，每一对 RGB 图像均发生变化。对于 100 个提示词，跨 GPU 的单张图像 MAE/SSIM/LPIPS 为 5.656/0.913/0.050（列出的 200 对、100 个唯一对）；B8 为 4.944/0.938/0.032（列出的 16 对、8 个唯一对，其中包含一个构图和一次精确重放）。记录的来源、权重、软件栈、设置、初始潜变量和条件信息均相匹配。

对于 FLUX 3 Klein bf16，在 512×512 和 50 个步数下进行的独立匹配 RTX A6000/RTX 6000 Ada 对比中，跨 GPU 的单张图像 SSIM/LPIPS 为 0.699/0.139，B8 为 0.809/0.122。阶段100的单张图像池包含列出的 200 对、但实际为 100 个唯一对；B8 包含列出的 16 对、但实际为 8 个唯一对，其中包括重放。阶段10的结果未与这些数值合并。

图形处理器和批处理效应取决于模型和协议。Janus 的匹配结果处于同一量级；在固定输入的通义千问 P100 对比中，测得的跨 GPU 差异小于批处理效应。

F. 根因分析

批次形状变化：首次观察到的分歧操作分别是 Janus-Pro-7B bf16 中的第11层均方根归一化，以及 SDXL 中的下采样器

表 IX 提示词类型细分。阴影标示同表 V。

提示词类别	发现
高频图案 结构相似性指数	
数字、文本、计数器	明显更差，数字错误
跨模型迁移	依赖模型的排名

卷积，以及通义千问 fp8 中共享的激活量化尺度。固定双行重放消除了 Janus 和 SDXL 中的局部差异，并缩小了通义千问中的差异，但没有任何单一干预能恢复完整流水线。在 Janus 中，组合归一化、线性层和注意力修复消除了所测试 bf16 跟踪中的全部 576 个词元不匹配；INT8归一化干预无法推广到所测试的五个提示词。通义千问的尺度干预将平均绝对误差从 2.591 降至 1.390，并从 20.240 降至 15.880，但未能恢复单独运行时的一致性。批次形状会改变特定操作，而修复其中一个操作不足以使整个流水线具备批次不变性。

在 H100 NVL 上单独复现的 Janus BF16 RMSNorm 涵盖 62 个位置（位置 2 至 63），每个位置包含 8192 个元素。捕获的 A6000 批次 1 参考图像与 H100批次大小为1的输出在全部 62 个位置上均精确匹配。原生 H100 的批次大小 1 与批次大小 8 仅在位置 57 存在差异：8192 个元素中有 11 个不同，最大绝对差为 0.0009765625，均方根差为 1.8344e-05。固定两行的批次大小 8 重放以及 fp64批次大小1与批次大小8对照在全部 62 个位置上均精确匹配。fp64参考结果与原生批次大小 1 的输出在位置 16 有 18 个元素不同，在位置 57 有 11 个元素不同，最大绝对差相同。此结果仅限于单独复现 RMSNorm，不代表完整模型结论。

G. 提示词类型细分

提示词类别。 接近分辨率极限的高频图案（如棋盘格、光栅和倾斜条纹）包含了许多结构相似性指数最低的案例。具体哪些提示词会出问题取决于模型，因此逐提示词排名无法跨模型沿用。包含数字、文本和计数器的提示词构成一个独特类别：如第四节 IV-C 所示，批处理图像可能在语义上出错，因此影响更严重，而不仅仅是有所不同。表 IX 总结了这一细分情况。

V. 讨论

影响。 这些结果直接影响文生图模型的评估和运行方式。首先，随机种子并不能保证结果可复现。仅报告提示词和随机种子并不足够；批次大小、精度和图形处理器也会决定输出，因此论文和基准测试应报告服务配置。其次，混合不同批次大小、精度或硬件进行的基准测试和图像质量比较可能产生偏差，因为测得差异的一部分源自执行设置，而非模型本身。第三，对生成图像进行审计、A/B 测试和安全性评估时也需要同样谨慎：即便模型本身没有任何变化，服务栈的更改也可能改变生成图像。

可能的原因。该模式与批处理依赖的内核执行一致：归约顺序会随张量形状而变化，而 8 位格式会将微小的数值差异放大为可见差异，fp8 动态量化还会引入一个批次级缩放因子，使批处理中的请求彼此耦合。要证实这一解释，需要进行内核级插桩，这将留待未来工作。

自回归模型为何更敏感。我们测得的两种自回归模型表现出最大的影响（Emu3.5 的平均结构相似性指数分别为 0.580 和 0.551，Janus-Pro-7B 的分别为 0.261 和 0.229 ），而对于这两种模型，重放组都能精确匹配批处理图像，因此差异由批处理决定，而非采样。一个可能的解释是自回归解码的序列结构。图像由一长串离散词元生成，每个词元都以之前的所有词元为条件，因此，微小的数值变化一旦导致某个词元翻转，就会改变后续每个词元的上下文，并使扰动沿序列累积。相比之下，扩散模型在每一步都会并行更新整个潜变量，因此能够平均或收缩微小扰动，而不是将其传播下去。两项观察支持这一解释。首先，Janus-Pro-7B 在贪心解码和采样解码下的影响相近，这与近似持平的 logit 发生翻转相符，而非随机抽样所致。其次，我们测得的扩散配置均未达到自回归模型的影响幅度。Janus 跟踪证据仅限于经过测试的 bf16 跟踪和提示词，因此这一解释仍属假设。

对量化的依赖。位宽本身并不能决定其影响。在我们测量的所有阶段，int8 都使 HiDream-O1 保持位不变，并使 FLUX 3 Klein 的变化低于可见程度（第 10 阶段的结构相似性指数为 0.998）；但它会使 Emu3.5 的表现低于 bf16（0.551 对比 0.580）。在我们测量的所有模型中，fp8 都是精度表现最差的：第 10 阶段，Qwen-Image 为 0.811，FLUX 3 Klein 为 0.829，HiDream-O1 为 0.748；到第 100 阶段则降至 0.686。一种合理的解释是各格式的执行方式不同。我们测试的 int8 配置只对权重进行量化，并反量化到 bf16 计算路径，因此算术内核和归约顺序与 bf16 运行时一致。fp8 则会改变计算精度本身：缩放和舍入在内核中执行，其归约顺序取决于批次形状，并且批次级缩放因子会使同一批处理中的请求相互耦合。按照这一解释，如果计算路径不变，批处理带来的影响就很小；而当批处理影响到算术运算时，其影响便会更大。Emu3.5 并不完全符合这一模式，因为尽管其 int8 配置与 bf16 共用 bf16 计算路径，批次敏感性却更高。除上述限定范围的跟踪和干预外，我们并未对完整的内核集进行插桩，因此这些解释仍是假设。

缓解方案。 对于重视可复现性的设置，我们的测量结果表明有三种选择。首先，在模型选择方面，我们矩阵中的扩散模型表现始终明显优于自回归模型：测得的扩散模型最低平均结构相似性指数为 0.686（HiDream-O1、fp8、第 100 阶段），而 Emu3.5 为 0.580 和 0.551，Janus-Pro-7B 为 0.261 和 0.229；不过，自回归模型的分数基于 10 个提示词的提示集，而扩散模型最多使用 100 个提示词。其次，使用批次1生成。对于我们测试的每个模型和精度，批次1重跑都能做到位精确（表 IV），因此批次1可以消除给定图形处理器上的批次

表 X 在提示词和随机种子固定时，所研究因素的影响。阴影标示同表 V：绿色表示无影响，黄色表示影响较小，橙色表示影响较大，红色表示影响最大。

因素	影响
批处理存在性（批处理与单独处理）	最大
模型家族（自回归与扩散）	最大
精度（8 位格式与 16 位格式）	较大
图形处理器（RTX 6000 Ada 与 RTX A6000，Janus bf16）	较大
提示词类型（高频图案）	大
批次大小超过 8	小
仅权重 int8 与 bf16 对比（HiDream-O1 和 FLUX 3 Klein 均无）	小
同一运行的精确重放	none
配套提示词和位置（已测试扩散模型和 Janus 的范围）	none

效应，但代价是吞吐量降低。它并不能让输出具备可移植性：对于 Janus-Pro-7B，在固定提示词和随机种子的情况下，RTX 6000 Ada 和 RTX A6000 上生成的 10 个提示词中，没有任何一个输出逐位相同。第三，按模型选择精度，因为不同精度的排序并不通用。在所有测得 fp8 的模型中，它都是最差的 8 位格式（Qwen-Image 为 0.811，FLUX 3Klein 为 0.829，HiDream-O1 在第 10 阶段为 0.748），应当避免使用。对于 SDXL，fp16 是最稳妥的选择，因为 bf16 和 int8 都会放大这种效应；对于 Qwen-Image 也是如此，在 8 个提示词中有 7 个，bf16 的批次敏感性高于 fp16。FLUX 3 Klein 和 HiDream-O1 能够以低于可见程度或零误差容忍 bf16 和 int8；Emu3.5 使用 bf16 时的性能下降（0.580）也小于使用 int8 时（0.551）。

局限性。本研究涵盖六种模型和三种 GPU 类型，大多数模型的每组提示词数最多为 100（SDXL 的评分提示词数为 25；bf16 除外，其学习感知图像块相似度和精确匹配评估最多覆盖 100 条提示词），且每个提示词集使用一个基础随机种子。我们测量了其影响，但未评估完整的服务修复方案；测试的局部干预并不能使整个流水线实现批次不变。图像模型的批次不变执行仍是一个尚待解决的问题。

VI. 结论

在 GPU 上，固定提示词和随机种子并不能固定图像。在六种扩散和自回归文生图模型中，批处理会改变几乎所有输出，而低精度通常会放大这种变化，有时甚至会生成肉眼可见地更差的图像。本文报告的重放控制均为位精确，因此这种效应是执行设置本身的属性，而非采样噪声所致。在我们的矩阵中，批处理是否存在的影响最广泛，模型家族决定了效应的规模，而 GPU 可能改变每一张图像；GPU 效应与批处理效应的相对大小取决于模型和实验方案。因此，若要实现可复现的文生图评估，除了提示词和随机种子，还必须报告批次大小、精度和 GPU。

致谢

本研究获得了 Agencia Estatal de Investigación (AEI)（doi:10.13039/501100011033）在 ANNA 资助项目（PID2025-169490OB-C21）下的支持，获得了由 CM（ORDEN 5696/2024）资助的 TUCAN6-CM（TEC-2024/COM460）的支持，并获得了欧盟地平线欧洲研究与创新计划项目 BRIDGE-AI（资助协议 101299050）的支持。

参考文献

[1] B. Atil、S. Aykent、A. Chittams、L. Fu、R. J. Passonneau、E. Radcliffe、G. R. Rajagopal、A. Sloan、T. Tudrej、F. Ture、Z. Wu、L. Xu 和 B. Baldwin，“‘ 确定性 ’ 大语言模型设置的非确定性 ”，2024 年。[2] J. Yuan 、H. Li、X. Ding、W. Xie、Y.-J. Li、W. Zhao、K. Wan、J. Shi、X. Hu 和 Z. Liu，“理解并缓解大语言模型推理中的非确定性的数值来源”，2025年。[3] H. He 和 Thinking Machines Lab，“攻克大语言模型推理中的非确定性”，https://thinkingmachines.ai/blog/defeating-nondeterminism-in-llm-inference/，2025年，访问日期：2026-11-11。[4] T. Fu、G. Mart´ınez、J. Conde、C. Arriaga、P. Reviriego、X. Qi 和 S. Liu，“超越可复现性：词元概率揭示大语言模型的非确定性”，*IEEE* 计算机汇刊，2026年，另见 arXiv:2601.06118。[5] LMSYS Org，“迈向 SGLang 中的确定性推理和可复现的强化学习训练”，https://www.lmsys.org/blog/ 2025-09-22-sglang-deterministic/，2025年，访问日期：2026-11-11。[6] vLLM Project，“批次不变性”，https://docs.vllm.ai/en/stable/ features/batch invariance/，2026年，访问日期：2026-11-11。[7] S. Zhao、J. Yang、Q. Chen、J. Hu、J. Zheng、G. Chen 和 X. Chen，“CoRun：填充对于确定性大语言模型推理而言简单且高效”，2026年。[8] X. Qi、P. Reviriego、J. Conde 和 S. Liu，“BlockLock：采用固定批次几何结构实现可复现的大语言模型推理”，2026年。[9] R. Gond、A. K. Kamath、R. Ramjee 和 A. Panwar，“LLM-42：利用经验证的推测实现大语言模型推理的确定性”，2026年。[10] P. Xie、X. Zhang 和 S. Chen，“RepDL：位级可复现的深度学习训练与推理”，2025年。[11] E. Kim、I. Isozaki、N. Sirkin 和 M. Robson，“生成式人工智能的可复现性与共识”，2023年。[12] Z. Fang、Z. Jiang、H. Chen、X. Zhang、T. Li、K. Tang、X. Li 和 J. Li，“数据集级指标削弱非确定性：对扩散语言模型非确定性的细粒度评估”，2026年。[13] Z. Wang、A. C. Bovik、H. R. Sheikh 和 E. P. Simoncelli，“图像质量评估：从误差可见性到结构相似性”，*IEEE*图像处理汇刊，第 13 卷，第 4 期，第 600–612 页，2004年4月。[14] R. Zhang、P. Isola、A. A. Efros、E. Shechtman 和 O. Wang，“深度特征作为感知度量的非凡有效性”，载于*IEEE*计算机视觉与模式识别会议（*CVPR*）论文集，2018年，第 586–595 页。[15] S. Shanmugavelu、M. Taillefumier、C. Culver、O. Hernandez、M. Coletti 和 A. Sedova，“浮点非结合性对高性能计算和深度学习应用可复现性的影响”，2024年。[16] PyTorch，“可复现性”，https://docs.pytorch.org/docs/main/notes/randomness.html，2026年，访问日期：2026-11-11。[17] NVIDIA，“不同 GPU 架构之间的确定性？ ”，GitHub issue，framework-reproducibility代码库，https://github.com/NVIDIA/ framework-reproducibility/issues/28，2023年，访问日期：2026-11-11。[18] O. E. Gundersen、K. Coakley、C. Kirkpatrick 和 Y. Gil，“机器学习中不可复现性的来源：综述”，2022年。[19] C. Summers 和 M. J. Dinneen，“神经网络优化中的非确定性与不稳定性”，载于第 *38* 届国际机器学习会议论文集，《机器学习研究论文集》系列，第 139 卷，2021年，第 9913–9922 页。[20] S. Woo，“GeulPUv2”，https://github.com/wwoosshh/GeulPUv2，2026年，访问日期：2026-11-11。[21] Z. Zhou、P. Yang、Y. Liang、M. Liu 和 J. Lu，“SynerDiff：用于快速并行扩散模型推理的协同连续批处理”，2026年。[22] D. Podell、Z. English、K. Lacey、A. Blattmann、T. Dockhorn、J. M¨uller、J. Penna 和 R. Rombach，“SDXL：改进用于高分辨率图像合成的潜在扩散模型”，载于国际学习表征会议，2024年。[23] P. Esser、S. Kulal、A. Blattmann、R. Entezari、J. M¨uller、H. Saini、Y. Levi、D. Lorenz、A. Sauer、F. Boesel、D. Podell、T. Dockhorn、Z. English、K. Lacey、A. Goodwin、Y. Marek 和 R. Rombach，“缩放用于高分辨率图像合成的整流流变换器”，载于国际机器学习会议，2024年。[24] C. Wu、J. Li、J. Zhou、J. Lin、K. Gao、K. Yan、S.-m. Yin、S. Bai、X. Xu、Y. Chen、Y. Chen、Z. Tang、Z. Zhang、Z. Wang、A. Yang、B. Yu、C. Cheng、D. Liu、D. Li、H. Zhang、H. Meng、H. Wei、J. Ni、K. Chen、K. Cao、L. Peng、L. Qu、M. Wu、P. Wang、S. Yu、T. Wen、W. Feng、X. Xu、Y. Wang、Y. Zhang、Y. Zhu、Y. Wu、Y. Cai 和 Z. Liu，“Qwen-Image 技术报告

,” 2025年。

[25] Q. Cai, J. Chen, Y. Chen, Y. Li, F. Long, Y. Pan, Z. Qiu, Y. Zhang, F. Gao, P. Xu, Y. Wang, K. Yu, W. Chen, Z. Feng, Z. Gong, J. Pan, Y. Peng, R. Tian, S. Wang, B. Zhao, T. Yao, and T. Mei, “Hidream-i1：一种具有稀疏扩散 Transformer 的高效图像生成基础模型”，2025年。[26] HiDream.ai, “Hidream-o1-image：一种采用像素级统一 Transformer 的原生统一图像生成基础模型”，2026年。[27] Black Forest Labs, “Flux 3 image klein：使用开放权重进行图像编辑”，https://bfl.ai/models/flux-3-image-klein，2026年，访问日期：2026-11-11。[28] X. Chen, Z. Wu, X. Liu, Z. Pan, W. Liu, Z. Xie, X. Yu, and C. Ruan, “Janus-pro：通过数据和模型扩展实现统一的多模态理解与生成”，2025年。[29] Y. Cui, H. Chen, H. Deng, X. Huang, X. Li, J. Liu, Y. Liu, Z. Luo, J. Wang, W. Wang, Y. Wang, C. Wang, F. Zhang, Y. Zhao, T. Pan, X. Li, Z. Hao, W. Ma, Z. Chen, Y. Ao, T. Huang, Z. Wang, and X. Wang, “Emu3.5：原生多模态模型是世界学习者”，2025年。[30] J. Han, J. Liu, Y. Jiang, B. Yan, Y. Zhang, Z. Yuan, B. Peng, and X. Liu, “Infinity：面向高分辨率图像合成的逐位自回归建模扩展”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2025年，第15 733–15 744页。[31] X. Li, Y. Liu, L. Lian, H. Yang, Z. Dong, D. Kang, S. Zhang, and K. Keutzer, “Q-diffusion：扩散模型量化”，2023年。[32] C. Chen, C. Giannoula, and A. Moshovos, “用于高效高质量扩散模型的低位宽浮点量化”，2024年。

Javier Conde 在西班牙马德里的马德里理工大学获得远程信息处理博士学位。目前，他是该校副教授。其研究兴趣包括数字孪生、大数据、语义网和人工智能。

Xiuyuan Qi 于2022年获得中国哈尔滨哈尔滨工程大学电子信息工程学士学位。目前，他正在中国成都电子科技大学攻读博士学位。其当前研究兴趣包括低功耗人工智能处理器的容错设计和计算机体系结构设计。

Shanshan Liu 于2018年获得中国哈尔滨工业大学微电子学与固体电子学博士学位。2018年至2021年，她在美国波士顿东北大学担任博士后研究员；2021年至2023年，她在美国拉斯克鲁塞斯新墨西哥州立大学担任助理教授。她现任电子科技大学信息与通信工程学院教授。

PedroReviriego 于1994年和1997年分别获得马德里理工大学（西班牙马德里）的电信工程硕士和博士学位。1997年至2000年，他在马德里的Teldat担任研发工程师。2000年，他加入Massana；2004年至2007年，他在LSI公司担任杰出技术人员。2007年至2018年，他任职于安东尼奥·德·奈布里哈大学，2018年至2022年任职于马德里卡洛斯三世大学。他目前任职于马德里理工大学。

附录 A 模型检查点与工具

表 A1 列出了所用的精确检查点，表 A2 列出了主要软件工具。FLUX 3 Klein 基础版仓库设有访问限制；下载前已接受许可证。所有链接均于 2026-11-11 验证。

附录 B 实验详情

本附录介绍通用实验协议和模型特定设置。附录 C 提供了更多视觉对比。

通用协议。对于每个模型、精度和提示词：A0 和 A1 实验组以批次大小 1 单独生成图像；B 实验组则在包含 8 个提示词的批处理中生成图像，并按轮换顺序混合提示词；对于 SDXL、Qwen-Image-2.1 和 HiDream-O1，C 实验组逐字重放第 0 轮批处理；FLUX 3 Klein、Emu3.5 和 Janus-Pro-7B 使用 R 实验组重放 B 实验组。批次 1 参考图像为 A0（r0）。提示词分为 S 组，每组 P 个提示词；每组使用一个基础随机种子，各组使用不同的随机种子。指标：逐位相同的比例（解码后的像素数组相等）、0 到 255 范围内的平均绝对误差、结构相似性指数（scikit-image 0.26.0，RGB，原生分辨率，数据范围 255，通道平均值）以及学习感知图像块相似度（lpips 0.1.4，AlexNet 骨干网络，输入缩放至 $[-1, 1]$）。完整研究分三个嵌套阶段进行（每组$P = 10$、25 和 100 个提示词，每组一个随机种子）；每个阶段都是一个检查点，其结果均为最终运行结果的有效子集。运行时间和内存用量见附录 D。

SDXL base 1.0。单一配置下使用 fp16 和 bf16 精度； 1024×1024, 30 步，引导强度 7.5； $S = 3$, 在嵌套阶段中分别为 25 和 100；B 实验组 = 256（阶段 10）生成 632 张（阶段 25）和 2504 张（阶段 100）图像，C 组 = 16；A臂对每组 P 张图像进行 3 次完全相同的单独重复运行（r0、r1、r2）（阶段 10 为 30 张图像，阶段 25 为 75 张，阶段 100 为 300 张）；每次运行生成的 PNG 图像：阶段 10 为 302 张，阶段 25 为 723 张，阶段 100 为 2820 张。图形处理器：RTX A6000。软件：torch 2.6.0（CUDA 12.4）、diffusers 0.31.0、transformers 4.46.3、NumPy 1.26.4。阶段 2 的运行设置 PyTorch 分配器标志 expandable segments:True，该设置仅改变内存分配，不影响数值结果。完整运行时间：1.22 小时（M）。

通义千问-Image-2.1。精度为 bf16 和 fp8； $S = 2$; 768×768用于所选配对研究；历史 H100 阶段序列和已完成的 B32 结果见 512×512 ，并单独报告；40 步（流水线默认值），关闭引导强度（true cfg scale 为 1.0），基础随机种子为 20261110；B 实验组在每个阶段 = 64 （8 轮，每轮 8 个），C 组 = 8，A0 和 A1 组各持有P 张图像；每次运行生成的 PNG 图像：阶段 10 为 92 张，阶段 25 为 122 张，阶段 100 为 272 张，另有阶段 100 的 B24 批次梯度组 416 张（B = 192，C = 24）；阶段 100 的 B16 批次梯度组在每种精度下各持有 344 张生成的 PNG 图像（A= 200, B = 128，C = 16），B32 组在 fp8 下持有 488 张（A = 200，B = 256，C = 32）。此处报告的已完成 bf16 B32 结果来自一次独立的 512×512 H100 运行，包含针对 100 个提示词的 256 组批处理与单独运行配对；它并非 768 像素方案的延续。图形处理器：H100 NVL。软件：torch 2.14.1、diffusers 0.41.0、transformers 5.20.2、torchao 0.18.0。

FLUX 3 Klein 基础版。精度为 bf16、仅权重 int8（torchao）和 fp8（torchao float8）； 512×512，50 步，引导强度为 3.5； $S = 3$、 $P = 10$、25 和 100，遍历嵌套阶段；R 组替代 C 组并重放 B 组；三种精度下的阶段尺寸相同：阶段 10 时 A0 = 10、A1 = 10、B = 64、R= 64 （每种精度生成 148 张 PNG 图像），阶段 25 时 A0 =25、A1 = 25、B = 96、R = 96 （242 张），阶段 100 时 A0 =100、A1 = 100、B = 344、R = 344 （888 张）；已配置批次梯度 8、16、32；每个单元重复 2 到 3 次，配置完全相同。图形处理器：H100 NVL（多实例 GPU 3g.47gb）。软件：torch 2.15.0（CUDA 13.2）、diffusers 0.42.1、transformers 5.22.2、torchao 0.18.0、NumPy 2.4.6。

HiDream-O1-Image。精度：bf16、int8（通过 torchao Int8WeightOnly 对语言模型线性层进行处理）和 fp8；512×512, 50 步，引导强度 5.0； $S = 3$、 $P = 10$, 在嵌套阶段中分别为 25 和 100；A0 和 A1 组各包含 P 张图像；批处理大小为 8 时，每个阶段的 B = 128 和 C = 8 （第 10 阶段生成 156 张 PNG 图像，第 25 阶段生成 186 张，第 100 阶段生成 336 张）；第 100 阶段的批次梯度还包括批处理大小 16（B = 256，C = 16，472 张）、批处理大小 24（B = 384，C = 24；生成 608 张 PNG 图像，与 A0 的 508 对评分配对）以及批处理大小 32（B = 512，C = 32，744 张）；每个单元使用相同配置重复 2 到 3 次。图形处理器：H100 NVL（多实例 GPU 3g.47gb）。软件：torch 2.14.1（CUDA 13.0）、diffusers 0.39.0、transformers 4.57.1、torchao 0.18.0。原始 HiDream-O1 生成环境未记录 Python 或 NumPy 版本，因此不予报告。较新的匹配源代码环境是独立环境，不能替代历史配置。

Emu3.5-Image。34B 自回归模型，采用贪心 argmax 解码（不使用采样种子）；语言模型堆栈使用 bf16 计算，另有一个 int8 变体，对 448 个量化线性层使用 torchao 权重专用量化；$P = 10$, $S = 1$; A0、A1、B（每组 8 个单元）和 R 组各包含 10 张图像，在第 10 阶段每次运行生成 40 张 PNG 图像；在 bf16 和 int8 下，第 100 阶段各包含 100 张图像，每次运行共 400 张；批次大小为 8，引导强度为 2.0， 32×32 令牌网格（512×512 张图像），随机种子为 20261110，使用即时注意力。图形处理器：H100 NVL 完整设备（bf16 需要超过 48 GB）。软件：torch 2.14.1（CUDA 12.6）、Python 3.10。int8 重复运行可精确复现结果。

Janus-Pro-7B。自回归模型； 384×384张图像，每张包含 576 个图像令牌； $P = 10$ （与 Emu3.5 使用相同的提示词）， $S = 1$；采用贪心（argmax）和采样（$T = 1.0$，cfg 5.0）解码；批处理分支使用 2 个单元（8+2）；评分时，贪心分支中的 G0/G1/GB/GR 分别对应 A0/A1/B/R；每个分支在每种解码模式下生成 10 张图像，即第 10 阶段每次运行生成 80 张 PNG 图像（每种模式 40 张）；使用 100 个提示词时，每种模式生成 100 张图像，每次运行共生成 800 张。第 10 阶段使用一块 RTX 6000 Ada，以 bf16 精度和 SDPA 注意力运行；随机种子为 20261110。跨 GPU 分支使用一块 RTX A6000；使用 bf16 精度运行的 100 个提示词实验均在两块 GPU 上完成。

附录 C 附加配置比较

图 A1–A5 对比了选定的批处理输出及其对应的批大小为 1 的参考结果。对于每种存在相关差异的额外配置，预留一个占位槽（布局与图 3 至图 8 相同：参考图像、SSIM 最差的图像、

表 A1 模型检查点（Hugging Face 仓库）。

模型	仓库
SDXL base 1.0	https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0
Qwen-Image-2.1	https://huggingface.co/Qwen/Qwen-Image-2.1
FLUX 3 Klein 基础版	https://huggingface.co/black-forest-labs/flux-3-图像-klein-base
HiDream-O1-Image	https://huggingface.co/HiDream-ai/HiDream-O1-Image
Emu3.5-Image	https://huggingface.co/BAAI/Emu3.5-Image
Janus-Pro-7B	https://huggingface.co/deepseek-ai/Janus-Pro-7B

表 A2
主要软件工具。

Tool	代码库
PyTorch	https://pytorch.org
diffusers	https://github.com/huggingface/diffusers
transformers	https://github.com/huggingface/transformers
torchao	https://github.com/pytorch/ao
scikit-image	https://scikit-image.org
学习感知图像块相似度	https://github.com/richzhang/PerceptualSimilarity

LPIPS 最差的图像对中的参考图像、LPIPS 最差的图像）。输出完全相同（SSIM $= 1$，LPIPS $= 0$）的配置（如采用 bf16 和 int8 的 HiDream-O1）不分配图像槽，仅显示表格行。

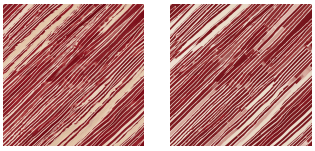


图 A1。SDXL，bf16，RTX A6000，批处理大小 8 与批处理大小 1 对比（数据类型探测，8 个提示词）。按平均绝对差计算最差的一对（提示词 02）：MAE 24.10（0–255）；SSIM 0.685，LPIPS 0.119。

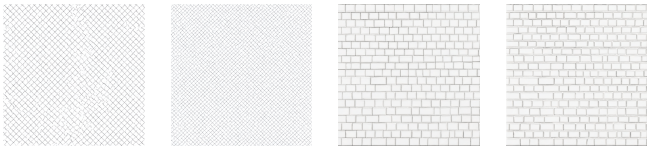


图 A2。Qwen-Image-2.1，fp8，H100 NVL。SSIM 最差的一对（提示词 09，精细网格）：SSIM 0.135，LPIPS 0.083，MAE 20.9；LPIPS 最差的一对（提示词 06）：LPIPS 0.152，SSIM 0.685，MAE 9.58。

图 A3。FLUX 3 Klein，bf16，H100 NVL。SSIM 最差的一对（提示词 09）：SSIM 0.989，LPIPS 0.002，MAE 2.37；LPIPS 最差的一对也是同一对（提示词 09，并列）。

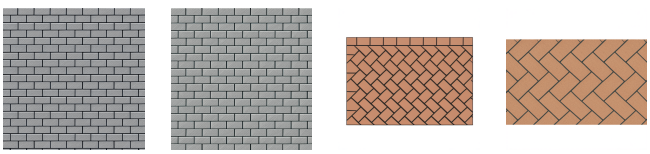


图 A4. Emu3.5，bf16，H100 NVL，第 10 阶段。最差的 SSIM 配对（提示词 5，灰色砖墙）：SSIM 为 0.216；最差的 LPIPS 配对（提示词 7）：LPIPS 为 0.496。

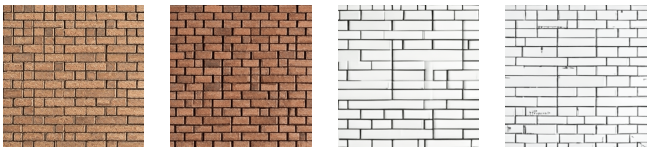


图 A5. Janus-Pro-7B，bf16，RTX 6000 Ada，采样解码。最差的 SSIM 配对（提示词 7，砖墙）：SSIM 为 0.127，LPIPS 为 0.320；最差的 LPIPS 配对（提示词 6，精细网格）：LPIPS 为 0.474，SSIM 为 0.259。

附录 D 执行细节

本附录详细介绍了在不同硬件平台上运行各模型实验所需的内存和计算时间。

表 A3 运行时间和内存使用量概览。除非标注为秒，否则时间单位均为小时，分别对应第 10、25 和 100 阶段。内存为报告的峰值；采样得到的 nvidia-smi 数值为下限。

模型、精度和图形处理器	运行时间；峰值内存
FLUX 3 Klein，bf16，H100-M	0.32/0.34/1.24 小时；33.4 GB
FLUX 3 Klein、int8、H100-M	0.40/0.44/1.60 小时；27.8 GB
FLUX 3 Klein，fp8，H100-M	0.39/0.42/1.54 小时；27.8 GB
HiDream-O1，bf16，H100-M	0.12/0.14/0.28 小时；18.2 GB
HiDream-O1，int8，H100-M	0.17/0.23/0.55 小时；11.2 GB
HiDream-O1、fp8、H100-M	0.17/0.25/0.63 小时；11.5 GB
SDXL、fp16、RTX A6000	0.54/1.28/5.03 小时；已分配 39.325 GB
SDXL，bf16，RTX A6000	0.531/1.264/4.74 小时；已分配 24.26 GB
Qwen-Image-2.1，bf16，H100 NVL	0.06/0.08/0.17 小时；已分配 46.198 GB
Qwen-Image-2.1，fp8，H100 NVL	0.07/0.10/0.24 小时；已分配 39.083 GB
Emu3.5，bf16，H100	0.322/0.762/2.962 小时；采样 92.978 GB
Emu3.5，int8，H100	0.938/2.271/8.85 小时；采样 81.926 GB