PR #21041 完整报告

sgl-project/sglang

[diffusion] model: Fix FLUX.1 output correctness

合并时间：2026-03-24 20:17

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21041

执行摘要

- 一句话：修复 FLUX.1 扩散模型的图像序列长度计算错误，消除块状图像并支持 2048x2048 分辨率生成。
- 推荐动作：建议开发者阅读此 PR，了解扩散模型中图像序列长度计算的关键细节，特别是 patchification 处理的正确方式。对于扩散模型管道的维护者，此变更值得关注以确保模型正确性，并可参考讨论中提到的与 diffusers 的差异进行进一步优化。

功能与动机

根据 PR body，FLUX.1-dev 存在两个正确性问题：导致输出图像出现块状瑕疵，并且无法生成模型本应支持的 2048x2048 分辨率图像。此 PR 旨在修复这些问题，提升输出质量，具体引用原文："Currently FLUX.1-dev has two correctness issues that deteriorate the output quality."

实现拆解

实现集中在单个文件 python/sglang/multimodal_gen/runtime/pipelines/flux.py 的 prepare_mu 函数中。关键改动是将 image_seq_len 的计算从 (int(height) // vae_scale_factor) * (int(width) // vae_scale_factor) 修改为 (int(height) // (vae_scale_factor * 2)) * (int(width) // (vae_scale_factor * 2))，以正确考虑 patchification 对序列长度的影响。第二个修改（self-attention 相关）已在 PR #20679 中修复，因此本 PR 仅包含此计算修复。

关键文件：

- python/sglang/multimodal_gen/runtime/pipelines/flux.py（模块 multimodal_gen/runtime/pipelines）：这是唯一修改的文件，包含 prepare_mu 函数的关键计算修复，直接影响 FLUX.1 扩散模型的输出正确性。

关键符号：prepare_mu

评论区精华

在 Issue 评论中，主要讨论包括：作者 avjves 指出第二个修改已在 PR #20679 中修复，因此本 PR 只包含第一个修改；mickqian 要求提供 diffusers 的输出对比，avjves 回复显示修复后质量改善（无块状图像、支持高分辨率），但与 diffusers 输出仍有轻微差异，推测是其他实现差异所致。讨论结论是修复显著改善了图像质量，支持了更高分辨率，但建议进一步验证与

diffusers 的对齐。

- 修复正确性与 diffusers 差异 (correctness): 修复显著改善了输出质量和分辨率支持，但与 diffusers 的差异需进一步验证。

风险与影响

- 风险：风险较低，变更局限于 prepare_mu 函数的计算逻辑，基准测试显示性能影响微小（大多数阶段延迟变化在±3% 以内）。潜在风险包括：核心扩散管道路径变更可能引入回归，需确保计算正确性；与 diffusers 输出的差异（avjves 提到 "still some discrepancies"）可能表明仍有未解决的正确性问题，需要额外测试验证。
- 影响：直接影响使用 FLUX.1 扩散模型的用户：输出图像质量从块状改善为清晰，并能够生成 2048x2048 分辨率图像，提升功能完整性和用户体验。对系统性能影响可忽略，基准测试中延迟变化最小（如 DenoisingStage 仅 +1.3%）。此修复增强了模型的多模态生成能力，与近期扩散模型修复 PR（如 #21042）形成协同。
- 风险标记：核心路径变更，潜在输出差异

关联脉络

- PR #20679 [diffusion] Fix self-attention replication for FLUX.1: 本 PR 中提到的第二个修改已在此 PR 修复，与本 PR 共同解决 FLUX.1 的正确性问题。
- PR #21042 [diffusion] fix Z-Image SP sharding for portrait and padded resolutions: 同属扩散模型修复 PR，涉及类似领域（扩散管道正确性），可关联分析。
- PR #21250 [diffusion] Fix torch.zeros typo in causal wan: 同属扩散模型修复 PR，关注正确性 bugfix，反映近期扩散模块的维护趋势。