PR #20679 完整报告

sgl-project/sglang

[diffusion] fix: fix accuracy for some image models

合并时间：2026-03-22 15:11

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/20679

执行摘要

本 PR 修复了 Qwen-Image、Qwen-Image-Edit 和 Z-Image 等多个图像扩散模型的准确性偏差问题，通过对齐官方 diffusers 实现、修正序列并行（SP）处理逻辑，确保单 GPU 与多 GPU 输出一致。变更影响扩散管道配置和注意力层，提升了模型生成质量，是扩散模块的重要改进。

功能与动机

动机源于这些模型在特定场景下输出与官方 diffusers 库不一致，尤其是在使用负提示、CFG（Classifier-Free Guidance）和 SP 并行时。PR body 中明确指出需要 "align CFG with official diffusers"、"respect negative-image" 和 "fix accuracy"，目标是消除偏差，提供更可靠的图像生成。具体问题包括 Qwen-Image 的 CFG 未匹配官方 norm rescale、Qwen-Image-Edit 的负提示处理不当、旋转频率精度漂移，以及 Z-Image 在 SP 下的 tokenization 和 caption 分片错误。

实现拆解

实现按模型和模块拆解，关键代码逻辑如下：

模型 / 模块	关键改动	代码示例（简要）
Qwen-Image	引入 true_cfg_scale，重写 postprocess_cfg_noise 进行 norm 匹配	noise_pred * (cond_norm / noise_norm) in qwen_image.py
Qwen-Image-Edit	保持旋转频率为 fp32，添加 SP 分片函数 _shard_qwen_edit_img_cache_for_sp	img_cache = shard_rotary_emb_for_sp(...) in qwen_image.py
Z-Image	移除 caption token 分片，改为复制后缀，修复 RoPE 偏移	num_replicated_suffix 参数在 zimage.py 和 layer.py 中
通用层	扩展 USPAttention 支持 num_replicated_suffix，优化 CFG 逻辑	_forward_with_replicated_suffix in layer.py

模型 / 模块	关键改动	代码示例（简要）
基类	添加钩子如 get_latent_dtype 和 gather_noise_pred_for_sp	在 base.py 中定义默认实现

评论区精华

Review 讨论较少，仅有 gemini-code-assist[bot] 确认修改正确。Issue 评论中用户 Rockdu 提供了关键测试结果：

> Rockdu: "Thanks for fixing this SP precision issue, here are some quantized test results on our side for reference ... Worst-case min_cosine vs single_gpu_ref ... Z-Image-Turbo: model_output 0.5039 → 1.0000 🤔 Fixed (exact match)"

这显示修复显著提升了多 GPU 下的输出一致性，无争议点，团队认可修复效果。

风险与影响

风险分析：

- 回归风险：CFG 修改在 qwen_image.py 中可能影响其他继承基类的模型，需全面回归测试。
- SP 并行风险：zimage.py 中的 caption token 处理变更在复杂 SP 场景下可能引入新偏差，需验证多 GPU 对齐。
- 性能影响：保持 fp32 旋转频率可能轻微增加内存使用，但确保了精度。
- 兼容性：USPAttention 扩展需确保不破坏现有注意力后端。

影响评估：

- 用户影响：直接受益于更准确的图像生成，提升产品可靠性。
- 系统影响：需更新模型配置和测试套件，确保跨环境一致性。
- 团队影响：加强扩散模型维护流程，关注官方对齐和 SP 优化。

关联脉络

与近期 PR 的关联揭示扩散模块的持续演进：

- PR #20862（添加 FireRed-Image-Edit 模型）：共享扩散模型配置框架，显示团队在扩展模型支持。
- PR #21122（清理扩散 Triton 内核）：技术领域重叠，聚焦性能优化和代码现代化。 本 PR 作为准确性修复的关键一环，补全了扩散模型在 SP 并行下的短板，与这些 PR 共同推动生态系统成熟。