PR #21042 完整报告

sgl-project/sglang

[diffusion] fix Z-Image SP sharding for portrait and padded resolutions

合并时间：2026-03-24 10:15

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21042

执行摘要

- 一句话：修复 Z-Image 序列并行 sharding，支持肖像和填充分辨率，避免图像损坏。
- 推荐动作：建议技术管理者和工程师精读此 PR，特别是 _build_zimage_sp_plan 的设计决策和 denoising.py 中的 gather 逻辑修改，以理解序列并行中处理图像几何形状的技术权衡。关注风险点，确保测试充分覆盖新路径，并评估对其他管道的潜在影响。

功能与动机

Issue #21021 报告 Z-Image-Turbo 在启用序列并行（Ulysses/SP）时，对于某些分辨率（如 720x1280 和 720x720）会产生损坏图像，原因是当前实现会交换 H/W 以便 sharding 总是沿较大空间轴，破坏原生图像几何形状。PR body 指出需要修复以保持原生几何形状，确保去噪过程中图像正确性。

实现拆解

实现方案拆解如下：1) 在 pipeline 配置模块（zimage.py）中新增 _build_zimage_sp_plan 方法，根据令牌填充成本选择沿原生高度或宽度 shard，移除 swap_hw 逻辑；2) 修改基础配置（base.py 和 ltx_2.py）中的 gather_latents_for_sp 函数签名，添加 batch 参数以确保 Z-Image 特定路径兼容性；3) 在模型层（zimage.py）调整 patchify_and_embed 和 forward 函数，添加 image_seq_len_target 参数以支持对齐序列长度；4) 在 pipeline 阶段（denoising.py）更新 _postprocess_sp_latents 函数，使用 pipeline_config 的 gather 方法处理 Z-Image，避免固定维度假设；5) 在测试配置（testcase_configs.py）中添加新测试用例，覆盖肖像分辨率路径。

关键文件：

- python/sglang/multimodal_gen/configs/pipeline_configs/zimage.py（模块 pipeline 配置）：核心修改，实现新的 SP 计划逻辑（_build_zimage_sp_plan），移除 swap_hw 并选择 sharding 轴，直接影响 Z-Image 序列并行的正确性。
- python/sglang/multimodal_gen/runtime/pipelines_core/stages/denoising.py（模块 pipeline 阶段）：关键修改，更新 _postprocess_sp_latents 函数以处理 Z-Image 的 gather 逻辑，避免固定维度假设，确保轨迹张量正确重建。
- python/sglang/multimodal_gen/configs/pipeline_configs/base.py（模块 基础配置）：基础配置修改，统一 gather_latents_for_sp 函数签名（添加 batch 参数），影响所有依赖此接口的管道，需保持兼容性。

- python/sglang/multimodal_gen/test/server/testcase_configs.py（模块 测试配置）：添加新测试用例（zimage_image_t2i_2_gpus_non_square），覆盖肖像分辨率路径，但可能未完全验证所有新逻辑，需关注测试覆盖。

关键符号：_build_zimage_sp_plan, gather_latents_for_sp, patchify_and_embed, _postprocess_sp_latents

评论区精华

Review 中核心讨论点包括：1) 正确性问题：BBuf 指出 denoising.py 中的 gather 逻辑假设固定维度，对于 Z-Image 的 W-shard 路径可能导致错误，Ratish1 确认并修复，通过重用 pipeline_config 的 gather 方法（评论位于 denoising.py:818）；2) 设计权衡：BBuf 建议将 _ceil_to_multiple helper 函数移到 utils.py，Ratish1 认为应保持文件特定性，最终未移动（评论位于 zimage.py:722）；3) 假设验证：BBuf 询问 image_padding_len 是否总是非负，Ratish1 解释由于下限确保不会为负，已解决（评论位于 zimage.py:776）。讨论聚焦于正确性和设计，关键疑虑已处理。

- gather 逻辑的正确性适配 (correctness): Ratish1 确认并修复，通过重用 pipeline_config 的 gather 方法，确保 Z-Image 特定路径正确处理。
- helper 函数是否应移到 utils.py (design): 未移动，保持原文件，设计权衡偏向于局部封装。
- image_padding_len 的非负性假设 (correctness): Ratish1 解释由于 image_seq_len_target 被下限确保（通过 _ceil_to_multiple），image_padding_len 不会为负，假设安全。

风险与影响

- 风险：技术风险具体包括：1) 回归风险：修改了基础配置中的 gather_latents_for_sp 签名，可能影响依赖此函数的其他管道（如 video 或 image 管道），需确保兼容性；2) 性能风险：新的 SP 计划选择逻辑可能增加计算开销，尤其是在每请求构建计划时；3) 测试覆盖不足：Issue 评论中 BBuf 指出缺少针对肖像和填充分辨率路径的回归测试，当前测试用例可能未完全验证新逻辑（testcase_configs.py 仅添加了一个测试）；4) 安全风险：无直接安全影响，但错误 gather 可能导致数据损坏。
- 影响：影响范围分析：1) 用户影响：Z-Image 模型用户在多 GPU 配置下现在能正确生成各种分辨率的图像，提升用户体验和模型可靠性；2) 系统影响：改进序列并行在扩散模型中的正确性，增强系统稳定性和可扩展性；3) 团队影响：需关注测试覆盖，并可能对其他类似管道（如 LTX）有借鉴意义，促进序列并行最佳实践。
- 风险标记：核心路径变更，缺少测试覆盖，接口兼容性风险

关联脉络

- 暂无明显关联 PR