PR #22374 完整报告

sgl-project/sglang

[diffusion] fix: fix cache dit refresh none mask

合并时间：2026-04-09 11:58

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22374

执行摘要

本 PR 修复了扩散模型缓存 DIT 集成中当 mask 策略为 None 时的 bug，并优化了调度器预热过程以避免多进程图像读写冲突。变更涉及核心缓存逻辑和分布式处理，提升系统稳定性和性能，适合关注扩散模型和分布式部署的工程师精读。

功能与动机

动机是修复一个潜在 bug：在 cache_dit_integration.py 中，当 scm_preset 参数为 None 时，原代码仍会调用 cache_dit.steps_mask 生成 mask，可能导致 TypeError 或逻辑错误。通过将 mask 生成条件化，确保仅在 scm_preset 非 None 时生成 mask，否则设为 None，避免无效调用。

实现拆解

- 缓存 DIT 集成修复：在 cache_dit_integration.py 的 refresh_context_on_transformer 和 refresh_context_on_dual_transformer 函数中，引入局部变量（如 steps_computation_mask），仅在 scm_preset 不为 None 时调用 cache_dit.steps_mask，否则置为 None。python if scm_preset is not None: steps_computation_mask = cache_dit.steps_mask(mask_policy=scm_preset, total_steps=num_inference_steps) else: steps_computation_mask = None
- 调度器预热优化：在 scheduler.py 中，重构 prepare_server_warmup_reqs 方法，提取 _prepare_shared_warmup_image_path 私有方法，使用 broadcast_pyobj 同步图像路径，避免多进程同时读写文件。
- 新增单元测试：添加 test_cache_dit_integration.py，通过模拟 cache_dit 等依赖，测试 mask 生成和刷新逻辑，确保修复的正确性。

评论区精华

Review 评论为空，变更由作者直接合并，未经过外部讨论。这表明变更可能被视为紧急修复或逻辑简单，但缺乏同行评审可能增加潜在风险。

风险与影响

- 技术风险：缓存逻辑修改可能引入回归，影响扩散模型推理正确性；分布式同步逻辑增加复杂性，有死锁或性能开销风险。
- 影响范围：直接影响扩散模型服务用户，提升稳定性；系统层面优化预热过程，减少多 GPU 环境下的冲突；团队需更新测试以确保覆盖。

关联脉络

与历史 PR 21204（扩散模型 RL 后训练）相关，同属 diffusion 模块的缓存和调度改进；与 PR 22384（池大小逻辑提取）关联，都涉及调度器重构，反映团队持续优化系统模块化和性能的趋势。