

PR #28724 完整报告

sgl-project/sglang

[NPU][FIX CI] Set the warm-up mode to "request" for NPU diffusion tests

合并时间: 2026-06-20 06:14

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/28724>

执行摘要

- 一句话: NPU 扩散测试 warm-up 模式改为 request
- 推荐动作: 可快速合并。建议后续为 NPU 添加专用的 warm-up 选项或文档说明, 避免依赖通用配置。值得关注 PR #28184 的 warm-up 实现。

功能与动机

PR 描述提到新添加的 warm-up 配置参数 (参考 #28184) 导致 NPU 场景性能退化, 因此需
要将 warm-up 模式调整为更合适的 request 模式。

实现拆解

1. 修改配置变量: 在 `python/sglang/multimodal_gen/test/server/ascend/testcase_configs_npu.py` 文件中, 将 `EXTRAS_DISABLE_WARMUP` 列表的值从 `["--server-warmup", "false"]` 改为 `["--warmup-mode", "request"]`。
2. 影响所有用例: 由于所有 NPU 测试用例均引用 `EXTRAS_DISABLE_WARMUP` 作为 `server extras`, 因此该变更同时改变了 `flux T2I`、`Wan2.1 T2V` 等用例的 warm-up 行为。
3. 无其他改动: 该 PR 仅包含这一行修改, 未涉及其他文件或逻辑变更。

关键文件:

- `python/sglang/multimodal_gen/test/server/ascend/testcase_configs_npu.py` (模块 测试配置; 类别 test; 类型 test-coverage): 核心变更文件, 修改了 warm-up 配置常量, 影响所有 NPU 扩散测试用例。

关键符号: 未识别

评论区精华

无实质性 review 讨论。机器人评论仅自动更新了 review 内容, 未提出技术质疑; ping1jing2 直接批准。

- 无实质讨论 (other): PR 被直接批准。

风险与影响

- 风险：低风险。仅修改测试配置参数，且 request 模式是已实现的正式选项，不会影响生产环境路径。但若 NPU 端 warm-up request 模式本身存在 bug，可能导致测试启动失败或结果不可靠，需结合 CI 运行结果确认。
- 影响：影响范围仅限于 NPU 扩散测试（Ascend 平台）；其他平台（NVIDIA/AMD）不受影响。用户无感知，仅测试维护者需关注 CI 状态。
- 风险标记：依赖于新增参数是否稳定

关联脉络

- PR #28184 Add warmup-mode argument: 本 PR 依赖 #28184 新增的 --warmup-mode 参数，将 NPU 测试配置从旧参数迁移至新参数。