PR #21682 完整报告

sgl-project/sglang

[diffusion] CI: relax pr-test threshold

合并时间：2026-03-30 20:23

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21682

执行摘要

本次 PR 放松了扩散模型在 PR 测试中的性能阈值，通过调整 perf_baselines.json 中的基准值，减少 CI 假阳性失败，以配合性能控制责任转移至 nightly-ci 的新策略。

功能与动机

变更的动机源于性能监控策略的调整：根据 PR body 描述，性能控制和跟踪已转移到 nightly-ci（具体在 nightly-test-diffusion-comparison 中）。因此，需要放松 PR 测试的阈值，避免因性能波动导致不必要的 CI 失败，让 PR 测试更专注于功能验证而非性能监控。

实现拆解

变更仅涉及一个文件：python/sglang/multimodal_gen/test/server/perf_baselines.json。具体修改包括：

- 将 pr_test 下的指标阈值提高：
  - *e2e* 从 0.15 增加到 0.2
  - denoise_stage 从 0.1 增加到 0.2
  - non_denoise_stage 从 0.6 增加到 0.8
  - 调整 TimestepPreparationStage 的值从 47.26 到 422.21

这些改动使得性能测试的容错范围更宽，适应了新的 CI 策略。

评论区精华

Review 过程非常简单，只有一个自动评论：

> gemini-code-assist[bot] commented: “I have no feedback to provide.” 没有其他人工评论或讨论，变更直接通过，未引发技术争议或设计权衡。

风险与影响

风险分析：

- CI 测试阈值放松后，可能无法及时捕获性能回归，例如 *e2e* 阈值提高可能导致端到端性能下降未被 PR 测试检测。

- 依赖 nightly-ci 进行性能监控，如果 nightly-ci 运行不稳定或监控不足，可能影响整体质量保证。

影响分析：

- 对用户：无直接影响，因为这是内部 CI 配置。
- 对系统：PR 测试通过率可能提高，减少开发流程中的中断。
- 对团队：需要确保 nightly-ci 有效运行，并可能调整测试策略以平衡快速反馈和深度监控。

关联脉络

从历史 PR 看，此变更与近期 diffusion 模块的 CI 改进相关：

- PR #21653 修复了 diffusion 仪表板图表显示问题，同属 diffusion CI 优化脉络。
- PR #21625 通过使用离线量化检查点提升 CI 测试稳定性，共享了测试配置和性能基准主题。

这些 PR 共同反映了团队在强化 diffusion 模块测试和监控方面的持续努力。