PR #24551 完整报告

sgl-project/sglang

ci: bump test_mimo_models.py est_time 330 → 610

合并时间：2026-05-07 05:35

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24551

执行摘要

- 一句话：上调 MiMo 测试预计时长避免超时
- 推荐动作：该 PR 是必要的 CI 配置修复，可快速合并。建议合并后观察至少一轮 CI 运行，确认超时问题是否解决。若仍偶发超时，可考虑进一步上调或分包（将两个测试类拆分到不同 est_time 条目）。

功能与动机

PR #23811 新增了第二个测试类（TestMiMoV2，模型 XiaomiMiMo/MiMo-V2.5，TP=8 DP=2，运行 MMMU + GSM8K + EAGLE 推测解码），但未同步上调 est_time。该文件现在需要运行约 500-640 秒，而 est_time 仍为 330 秒，导致自动分区器持续使 stage-c-test-8-gpu-h200 的 shard 0 过载，频繁触发 30 分钟的 Run test 超时上限（例如运行 ID 25428444359、25411981650）。

实现拆解

1. 修改 test/registered/8-gpu-models/test_mimo_models.py 第 9 行：将 register_cuda_ci(est_time=330, suite="stage-c-test-8-gpu-h200") 中的 est_time 从 330 上调至 610。
2. 该调整不涉及任何逻辑代码变更，仅更新测试预估时间元数据，使 CI 自动分区器能正确分配 shard，避免超时。

关键文件：

- test/registered/8-gpu-models/test_mimo_models.py（模块 测试；类别 test；类型 test-coverage）：上调 est_time 参数，从 330 调整至 610，以匹配 PR #23811 新增测试类后的实际运行时长，避免 CI 分区超时。

关键符号：未识别

关键源码片段

test/registered/8-gpu-models/test_mimo_models.py

上调 est_time 参数，从 330 调整至 610，以匹配 PR #23811 新增测试类后的实际运行时长，避免 CI 分区超时。

```
# test/registered/8-gpu-models/test_mimo_models.py
```

```
from sglang.test.ci.ci_register import register_cuda_ci

# 关键变更：将 est_time 从 330 上调至 610，以匹配实际运行时长
register_cuda_ci(est_time=610, suite="stage-c-test-8-gpu-h200")

class TestMiMoV2Flash(GSM8KMixin, SpecDecodingMixin, DefaultServerBase):
    # ... 测试类定义不变
```

评论区精华

该 PR 无实质性 review 讨论（Kangyan-Zhou 直接批准，无评论）。从 PR body 和 commit message 可知，超时问题是明确的机器负载不均导致，解决方案为直接上调 est_time 参数。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。仅修改测试注册的预估时间元数据，不影响任何功能逻辑。若新估时仍低于实际运行时间（例如因环境差异导致运行时间超过 610 秒），则超时问题可能复现。建议在合并后观察 CI 运行情况，必要时再次上调。
- 影响：
  - CI 稳定性：直接影响 stage-c-test-8-gpu-h200 分区的测试成功率，从频繁超时转为预期正常。
  - 无用户 / 系统侧影响：仅涉及测试基础设施配置，不改变运行时行为、性能或 API。
  - 风险标记：CI 配置调整，无代码逻辑变更

关联脉络

- PR #23811 [未完全匹配，推测为新增大模型测试的 PR]：本 PR 直接原因：#23811 新增了 TestMiMoV2 测试类，但未同步上调 est_time，导致超时。