PR #24205 完整报告

sgl-project/sglang

[AMD] fix moriep unittest failure

合并时间：2026-05-01 14:36

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24205

执行摘要

- 一句话：禁用 SpecV2 以修复 AMD MoE RIP 测试
- 推荐动作：建议合并此临时修复以恢复 CI 稳定性，并跟进后续的 SpecV2 + moriep 兼容性修复。

功能与动机

SpecV2 在 main 分支默认启用（PR #21062），导致 moriep 单元测试出现回归错误（参见 CI 日志链接）。本 PR 通过禁用 SpecV2 恢复之前的行为，作为临时修复。

实现拆解

在 test/registered/amd/test_moriep_small.py 文件的 TestMTPwithTBONormal 和 TestMTPwithTBOLowLatency 两个测试类的 setUpClass 方法中，添加环境变量 SGLANG_ENABLE_SPEC_V2=false，以在启动服务器时禁用 SpecV2。仅新增 2 行代码，无其他修改。

关键文件：

- test/registered/amd/test_moriep_small.py（模块 测试；类别 test；类型 test-coverage）：唯一修改的文件，通过添加环境变量禁用 SpecV2 来修复测试失败。

关键符号：setUpClass

关键源码片段

test/registered/amd/test_moriep_small.py

唯一修改的文件，通过添加环境变量禁用 SpecV2 来修复测试失败。

```
# test/registered/amd/test_moriep_small.py
# 在 TestMTPwithTBONormal.setUpClass 中添加了第 403 行：
        env["MORI_SHMEM_MODE"] = "ISOLATION" # avoid out of symmetric heap memory
        env["SGLANG_ENABLE_SPEC_V2"] = "false" # 临时禁用 SpecV2 以修复回归错误

# 在 TestMTPwithTBOLowLatency.setUpClass 中也添加了类似行：
        env["SGLANG_USE_AITER"] = "1"
        env["SGLANG_MORI_DISPATCH_DTYPE"] = "bf16"
        env["SGLANG_MORI_NUM_MAX_DISPATCH_TOKENS_PER_RANK"] = "4096"
```

```
env["SGLANG_ENABLE_SPEC_V2"] = "false" # 临时禁用 SpecV2
env["MORI_SHMEM_MODE"] = "ISOLATION"
```

评论区精华

无 review 评论。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低，仅在测试环境中禁用 SpecV2，不影响生产环境。但该修复是临时性的，若后续跟进 PR 未及时合并，测试覆盖可能不完整。
- 影响：仅影响 AMD moriep 单元测试的执行，使得测试能够通过 CI。对系统其他部分无影响。
- 风险标记：临时修复，无生产环境影响

关联脉络

- PR #21062 Enable SpecV2 by default: 本 PR 的动机正是由于该 PR 默认启用了 SpecV2 导致了 moriep 测试失败。