PR #21735 完整报告

sgl-project/sglang

fix ut test_moe

合并时间：2026-04-04 12:57

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21735

执行摘要

本 PR 修复了 XPU 环境下 DeepSeek OCR 测试的内存泄漏和配置问题，通过添加内存清理函数、分离测试类和调整内存分数，显著提升 CI 测试稳定性，属于常规维护性修复，对团队开发效率有积极影响。

功能与动机

PR body 简洁说明 "fix XPU ut test_moe"，结合 Issue 评论中多次出现的 /rerun-failed-ci 命令，推断动机是解决 XPU 测试套件中 MOE 测试因内存不足或资源泄漏导致的频繁失败，以确保 CI 流水线可靠运行，减少维护中断。

实现拆解

改动涉及四个文件，按模块拆解如下：

文件路径	变更内容	关键代码逻辑
test/srt/run_suite.py	添加 test_deepseek_ocr_triton.py 到 XPU 测试套件	TestFile("xpu/test_deepseek_ocr_triton.py", 360)
test/srt/xpu/test_deepseek_ocr.py	新增 _cleanup_xpu_memory 方法，修复图像路径，改进 tearDownClass	gc.collect(); torch.xpu.empty_cache()
test/srt/xpu/test_deepseek_ocr_triton.py	新增独立测试类，避免 unittest 发现重复	继承自 TestDeepSeekOCR，覆盖 setUpClass
test/srt/xpu/test_intel_xpu_backend.py	添加 _cleanup_xpu_memory 函数，调整 mem_fraction_static 为 0.4	在装饰器中调用清理函数，优化内存使用

评论区精华

Review 中仅有 Fridge003 的批准，无实质性讨论评论。提交历史显示 16 次迭代，如提交消息所示：

- "fix OOM"：解决内存溢出问题。

- "fix hang of loading weight"：修复权重加载挂起。
- "Prevent duplicate unittest discovery"：避免测试重复发现。 这表明修复过程通过提交迭代解决了内存和配置问题，但无公开技术讨论。

风险与影响

风险：

1. 内存清理函数 _cleanup_xpu_memory 可能无法彻底释放 XPU 内存，尤其是在异常情况下。
2. 硬编码内存分数 0.4 在 test_intel_xpu_backend.py 中可能不适用于不同硬件配置，导致测试失败。
3. 新增测试文件可能增加 CI 执行时间或引入冗余测试用例。

影响：

- 对用户：无直接影响，因变更限于测试代码。
- 对系统：提升 CI 测试在 XPU 环境下的稳定性和通过率，减少假阳性失败。
- 对团队：降低 CI 维护成本，提高开发效率和测试反馈可靠性。

关联脉络

从同仓库近期历史 PR 分析中，PR 21280 和 21851 均涉及 DeepSeek 标签，可能与当前 PR 共享模型测试上下文。整体上，sglang 仓库近期频繁出现测试和 CI 修复（如 PR 22083、22081），表明团队正持续优化测试基础设施和跨平台兼容性，本 PR 是这一趋势的一部分，专注于 XPU 环境下的稳定性和资源管理。